

# Consistent Structure Estimation of Exponential-Family Random Graph Models With Additional Structure

Michael Schweinberger

## Abstract

We consider the challenging problem of statistical inference for exponential-family random graph models given one observation of a random graph with complex dependence (e.g., transitivity). To facilitate statistical inference, we endow random graphs with additional structure. The basic idea is that random graphs are composed of sub-graphs with complex dependence. We have shown elsewhere that when the composition of random graphs is known,  $M$ -estimators of canonical and curved exponential families with complex dependence are consistent. In practice, the composition is known in some applications, but is unknown in others. If the composition is unknown, the first and foremost question is whether it can be recovered. The main consistency results of the paper show that it is possible to do so as long as exponential families satisfy weak dependence and smoothness conditions. These results confirm that exponential-family random graph models with additional structure constitute a promising direction of statistical network analysis.

## 1 Introduction

Exponential-family random graph models [13, 41, 19, 38, 22] are models of network data, such as disease transmission networks, insurgent and terrorist networks, social networks, and the World Wide Web [26]. Such models are popular among network scientists [26], because network data are dependent data and exponential-family random graph models enable network scientists to model a wide range of dependencies in network data.

Exponential-family random graph models of dependent network data were pioneered by Frank and Strauss [13]. The models of Frank and Strauss [13] and more general models [41, 19, 38, 22] are discrete exponential families of densities with countable support  $\mathbb{X}$ —the set of possible graphs with  $n$  nodes and binary or non-binary, count-valued edges—of the form

$$p_{\boldsymbol{\eta}}(\mathbf{x}) = \exp(\langle \boldsymbol{\eta}, s(\mathbf{x}) \rangle - \psi(\boldsymbol{\eta})), \quad \mathbf{x} \in \mathbb{X}, \quad (1)$$

where  $\langle \boldsymbol{\eta}, s(\mathbf{x}) \rangle$  denotes the inner product of a vector of natural parameters  $\boldsymbol{\eta} \in \{\boldsymbol{\eta} \in \mathbb{R}^{\dim(\boldsymbol{\eta})} : \psi(\boldsymbol{\eta}) < \infty\}$  and a vector of sufficient statistics  $s : \mathbb{X} \mapsto \mathbb{R}^{\dim(\boldsymbol{\eta})}$  and  $\psi(\boldsymbol{\eta})$  ensures that  $\sum_{\mathbf{x}' \in \mathbb{X}} p_{\boldsymbol{\eta}}(\mathbf{x}') = 1$ .

In general, statistical inference for exponential-family random graph models is challenging [15, 32, 9, 36], because exponential-family random graph models induce complex dependence [e.g., transitivity, 26] and in most applications of exponential-family random graph models the number of observations is  $N = 1$ : e.g., epidemiologists may be unable to collect more than  $N = 1$  observation of a contact network through which infectious diseases spread (e.g., Ebola, HIV).

While statistical inference for exponential-family random graph models given  $N = 1$  observation of a random graph with complex dependence is problematic, we facilitate statistical inference by endowing exponential-family random graph models with additional structure. The basic idea is that, while  $N = 1$  random graph with complex dependence is observed, the random graph is composed of a large number of subgraphs with complex dependence. The fact that the complex dependence is restricted to subgraphs induces local dependence, which in turn induces weak dependence and hence facilitates consistency results. We have shown elsewhere [35] that when the composition of the random graph is known,  $M$ -estimators of canonical and curved exponential-family random graph models with local dependence are consistent under weak conditions. While the composition of the random graph is known in some applications [e.g., in multilevel networks, 24], it is unknown in others. If the composition of the random graph is unknown, the first and foremost question is whether it can be recovered with high probability given  $N = 1$  observation of a random graph with complex dependence. The main consistency results of the paper show that it is possible to do so as long as the data-generating exponential-family random graph model satisfies weak dependence and smoothness conditions. These consistency results cover a wide range of canonical and curved exponential-family random graph models, including models with transitive edge terms and geometrically weighted edgewise shared partner terms [19, 18]. These results confirm that—while exponential-family random graph models without additional structure are problematic [15, 32, 9, 36]—exponential-family random graph models with additional structure constitute a promising direction of statistical network analysis.

The paper is structured as follows. Section 2 introduces exponential-family random graph models with additional structure. Section 3 discusses the main consistency results. Section 4 presents simulation results. Section 5 proves the main consistency results.

**Other, related literature** It is worth noting that two broad classes of exponential-family random graph models can be distinguished based on the underlying dependence assumptions: one class focuses on exponential-family random graph models with independence assumptions [e.g., the  $\beta$ -model, 17, 12, 29, 42], while the other class focuses on exponential-family random graph models with dependence assumptions [13, 38, 19]. The independence assumptions of the first class of models are simplistic, because edges in real-world networks tend to depend on other edges [16]. The dependence assumptions of the second class of models are problematic, because many models—both canonical and curved exponential-family random graph models—allow edges to depend on many other edges: e.g., the conditional independence assumptions of Frank and Strauss [13] allow the conditional distribution of each edge variable to depend on  $2(n - 2)$  other edge variables. Such dependence assumptions can induce strong dependence, which in turn can give rise to model

degeneracy [15, 32, 9]. In its most general form, model degeneracy means that exponential-family random graph models place much probability mass on sufficient statistic vectors close to the boundary of the convex hull of the set of sufficient statistics vectors [32]. As a consequence, maximum likelihood estimators either do not exist at all or do exist but are hard to obtain by numerical maximization procedures [15, 28]. To replace the strong dependence assumptions of many exponential-family random graph models by weak dependence assumptions, Schweinberger and Handcock [33] advanced the notion of local dependence. Schweinberger and Stewart [35] proved that when the composition of the random graph is known,  $M$ -estimators of canonical and curved exponential-family random graph models with local dependence are consistent under weak conditions. We are here concerned with random graphs with unknown composition.

## 2 Exponential-family random graph models with additional structure

In general, statistical inference for exponential-family random graph models given  $N = 1$  observation of a random graph with complex dependence is challenging. We facilitate statistical inference by endowing exponential-family random graph models with additional structure that induces weak dependence and hence facilitates consistency results.

Throughout, we consider random graphs with a set of nodes  $\mathcal{A} = \{1, \dots, n\}$  and a set of edges  $\mathcal{E} \subseteq \mathcal{A} \times \mathcal{A}$ , where edges between pairs of nodes  $(i, j) \in \mathcal{A} \times \mathcal{A}$  are regarded as random variables  $X_{i,j}$  with countable sample spaces  $\mathbb{X}_{i,j}$ . We focus on undirected graphs without self-edges—i.e.,  $X_{i,i} = 0$  and  $X_{i,j} = X_{j,i}$  with probability 1—but extensions to directed random graphs are straightforward. We write  $\mathbf{X} = (X_{i,j})_{i < j}^n$  and  $\mathbb{X} = \times_{i < j}^n \mathbb{X}_{i,j}$ .

To facilitate statistical inference given  $N = 1$  observation of a random graph with complex dependence, we assume that the random graph is endowed with additional structure in the form of a partition of the set of nodes  $\mathcal{A}$  into  $K \geq 2$  subsets of nodes  $\mathcal{A}_1, \dots, \mathcal{A}_K$ , called neighborhoods. Since the number of observations is  $N = 1$ , it is important that the additional structure induces weak dependence, so that consistency results can be obtained. We induce weak dependence by restricting dependence to within-neighborhood subgraphs  $\mathbf{X}_{k,k} = (X_{i,j})_{i \in \mathcal{A}_k < j \in \mathcal{A}_k}$  ( $k = 1, \dots, K$ ). The resulting exponential families induce a form of local dependence defined as follows [33].

**Definition. Exponential families with local dependence.** *An exponential family of densities of the form (1) with countable support  $\mathbb{X}$  satisfies local dependence as long as its densities satisfy*

$$p_{\boldsymbol{\eta}}(\mathbf{x}) = \prod_{k=1}^K p_{\boldsymbol{\eta}}(\mathbf{x}_{k,k}) \prod_{l=1}^{k-1} \prod_{i \in \mathcal{A}_k, j \in \mathcal{A}_l} p_{\boldsymbol{\eta}}(x_{i,j}) \quad \text{for all } \mathbf{x} \in \mathbb{X}. \quad (2)$$

We give examples of canonical and curved exponential families with local dependence in Sections 2.1 and 2.2, respectively. We discuss the well-known, but restrictive special case

of stochastic block models in Section 2.3 and demonstrate the added value of exponential families with local dependence relative to stochastic block models in Section 2.4.

## 2.1 Example: canonical exponential families with local dependence

An example of canonical exponential families with local dependence and support  $\mathbb{X} = \{0, 1\}^{\binom{n}{2}}$  is given by exponential families with neighborhood-dependent edge and transitive edge terms of the form

$$p_{\boldsymbol{\eta}}(\mathbf{x}) \propto \exp \left( \sum_{k \leq l}^K \eta_{1,k,l} \sum_{i \in \mathcal{A}_k, j \in \mathcal{A}_l} x_{i,j} + \sum_{k=1}^K \eta_{2,k,k} s_{k,k}(\mathbf{x}) \right),$$

where

$$s_{k,k}(\mathbf{x}) = \sum_{i \in \mathcal{A}_k < j \in \mathcal{A}_k} x_{i,j} \mathbb{1}_{i,j}(\mathbf{x}).$$

Here,  $\mathbb{1}_{i,j}(\mathbf{x}) = 1$  if the number of shared partners of nodes  $i \in \mathcal{A}_k$  and  $j \in \mathcal{A}_k$  in neighborhood  $\mathcal{A}_k$  satisfies  $\sum_{h \in \mathcal{A}_k, h \neq i,j} x_{h,i} x_{h,j} > 0$  and  $\mathbb{1}_{i,j}(\mathbf{x}) = 0$  otherwise. If  $x_{i,j} \mathbb{1}_{i,j}(\mathbf{x}) = 1$ , the edge between nodes  $i$  and  $j$  is called transitive. We note that in recent work [22, 20, 23, 35] transitive edge terms have turned out to be attractive alternatives to the triangle terms which have been used since the classic work of Frank and Strauss [13] but which possess undesirable properties [15, 32, 9].

## 2.2 Example: curved exponential families with local dependence

An example of curved exponential families with local dependence and support  $\mathbb{X} = \{0, 1\}^{\binom{n}{2}}$  is given by exponential families with neighborhood-dependent edge and geometrically weighted edgewise shared partner terms of the form

$$p_{\boldsymbol{\eta}}(\mathbf{x}) \propto \exp \left( \sum_{k \leq l}^K \eta_{1,k,l} \sum_{i \in \mathcal{A}_k, j \in \mathcal{A}_l} x_{i,j} + \sum_{k=1}^K \sum_{t=1}^{|\mathcal{A}_k|-2} \eta_{2,k,k,t} s_{k,k,t}(\mathbf{x}) \right),$$

where

$$s_{k,k,t}(\mathbf{x}) = \sum_{i \in \mathcal{A}_k < j \in \mathcal{A}_k} x_{i,j} \mathbb{1}_{i,j,t}(\mathbf{x}).$$

Here,  $\mathbb{1}_{i,j,t}(\mathbf{x}) = 1$  if the number of shared partners of nodes  $i \in \mathcal{A}_k$  and  $j \in \mathcal{A}_k$  in neighborhood  $\mathcal{A}_k$  satisfies  $\sum_{h \in \mathcal{A}_k, h \neq i,j} x_{h,i} x_{h,j} = t$  and  $\mathbb{1}_{i,j,t}(\mathbf{x}) = 0$  otherwise. A curved exponential-family parameterization is given by

$$\begin{aligned} \eta_{1,k,l}(\boldsymbol{\theta}) &= \theta_{1,k,l} \\ \eta_{2,k,k,t}(\boldsymbol{\theta}) &= \theta_{2,k} \left\{ \theta_{3,k} \left[ 1 - \left( 1 - \frac{1}{\theta_{3,k}} \right)^t \right] \right\}, \quad \theta_{3,k} > \frac{1}{2}. \end{aligned} \tag{3}$$

Such terms are called geometrically weighted edgewise shared partner terms [19, 18], because the natural parameters  $\eta_{2,k,k,t}(\boldsymbol{\theta})$  are based on the geometric sequence  $(1 - 1/\theta_{3,k})^t$ ,  $t = 1, 2, \dots$ . It is worth noting that the corresponding geometric series converges as long as  $\theta_{3,k} > 1/2$  and that  $\theta_{3,k} \leq 1/2$  is problematic on probabilistic and statistical grounds [32, 35]. The parameterization is called a curved exponential-family parameterization, because the natural parameter vector  $\boldsymbol{\eta}(\boldsymbol{\theta})$  is a non-affine function of a lower-dimensional parameter vector  $\boldsymbol{\theta}$ ; see Remark 5 in Section 3.2. Last, but not least, note that in the special case  $\theta_{3,k} = 1$  ( $k = 1, \dots, K$ ) the curved exponential family reduces to the canonical exponential family described in Section 2.1.

## 2.3 Example: stochastic block models

A well-known, but restrictive special case of exponential families with local dependence and support  $\mathbb{X} = \{0, 1\}^{\binom{n}{2}}$  are stochastic block models [27]. Stochastic block models assume that all edge variables  $X_{i,j}$  are independent given the neighborhood structure, which implies that  $p_{\boldsymbol{\eta}}(\mathbf{x})$  can be written as

$$p_{\boldsymbol{\eta}}(\mathbf{x}) \propto \exp \left( \sum_{k \leq l}^K \eta_{1,k,l} \sum_{i \in \mathcal{A}_k, j \in \mathcal{A}_l} x_{i,j} \right),$$

where  $\eta_{1,k,l}$  is the log odds of the probability of an edge between nodes in neighborhoods  $\mathcal{A}_k$  and  $\mathcal{A}_l$ . While stochastic block models can be used to detect communities in networks, such models fail to capture a wide range of dependencies encountered in networks—such as transitivity [26]—and are therefore misspecified models: see, e.g., the discussion of Snijders [37].

## 2.4 Added value of exponential families with local dependence

The added value of exponential families with local dependence relative to stochastic block models is rooted in the ability to capture a wide range of dependencies. To demonstrate, consider exponential families with neighborhood-dependent edge and transitive edge terms as described in Section 2.1 and let  $s_{k,k}(\mathbf{x})$  be the number of transitive edges in neighborhood  $\mathcal{A}_k$ . By well-known exponential-family properties [7, Corollary 2.5, p. 37], the expected number of transitive edges in neighborhood  $\mathcal{A}_k$  satisfies

$$\mathbb{E}_{\eta_{1,k,k}, \eta_{2,k,k} > 0} s_{k,k}(\mathbf{X}) > \mathbb{E}_{\eta_{1,k,k}, \eta_{2,k,k} = 0} s_{k,k}(\mathbf{X}), \quad k = 1, \dots, K,$$

where  $\mathbb{E}_{\eta_{1,k,k}, \eta_{2,k,k}} s_{k,k}(\mathbf{X})$  denotes the expectation of  $s_{k,k}(\mathbf{X})$  and  $\eta_{1,k,k}$  and  $\eta_{2,k,k}$  denote the natural edge and transitive edge parameter of neighborhood  $\mathcal{A}_k$ , respectively. In other words, the expected number of transitive edges in neighborhood  $\mathcal{A}_k$  is greater under exponential families with local dependence with  $\eta_{2,k,k} > 0$  than under stochastic block models with  $\eta_{2,k,k} = 0$ , assuming that both have the same edge parameters  $\eta_{1,k,k}$  ( $k = 1, \dots, K$ ).

## 2.5 Notation

Throughout,  $\mathbb{E} f(\mathbf{X})$  denotes the expectation of a function  $f : \mathbb{X} \mapsto \mathbb{R}$  of a random graph with respect to exponential-family distributions  $\mathbb{P}$  admitting densities of the form (2). We write  $\mathbb{P} \equiv \mathbb{P}_{\boldsymbol{\eta}^*}$  and  $\mathbb{E} \equiv \mathbb{E}_{\boldsymbol{\eta}^*}$ , where  $\boldsymbol{\eta}^* \in \Xi \subseteq \text{int}(\mathbb{N})$  denotes the data-generating natural parameter vector and  $\Xi \subseteq \text{int}(\mathbb{N})$  denotes a subset of the interior  $\text{int}(\mathbb{N})$  of the natural parameter space  $\mathbb{N} = \{\boldsymbol{\eta} \in \mathbb{R}^{\dim(\boldsymbol{\eta})} : \psi(\boldsymbol{\eta}) < \infty\}$ . We assume that  $\boldsymbol{\eta} : \Theta \times \mathbb{Z} \mapsto \Xi$  is a function of  $(\boldsymbol{\theta}, \mathbf{z}) \in \Theta \times \mathbb{Z}$ , where

$$\Theta \times \mathbb{Z} = \{(\boldsymbol{\theta}, \mathbf{z}) \in \mathbb{R}^{\dim(\boldsymbol{\theta})} \times \{1, \dots, K\}^n : \psi(\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})) < \infty\}.$$

Here,  $\boldsymbol{\theta}$  is a vector of neighborhood-dependent parameters of dimension  $\dim(\boldsymbol{\theta}) \leq \dim(\boldsymbol{\eta})$  while  $\mathbf{z}$  is a vector of neighborhood memberships of nodes. Observe that the natural parameter vectors of the canonical and curved exponential families described in Sections 2.1 and 2.2 can be represented in this form. The data-generating values of  $(\boldsymbol{\theta}, \mathbf{z}) \in \Theta \times \mathbb{Z}$  are denoted by  $(\boldsymbol{\theta}^*, \mathbf{z}^*)$ . The  $\ell_1$ -,  $\ell_2$ -, and  $\ell_\infty$ -norm of vectors are denoted by  $\|\cdot\|_1$ ,  $\|\cdot\|_2$ , and  $\|\cdot\|_\infty$ , respectively. Uppercase letters  $A, B, C > 0$  denote unspecified constants, which may be recycled from line to line.

## 3 Consistent estimation of neighborhood structure

Exponential-family random graph models with neighborhood structure induce local dependence, which in turn induces weak dependence and hence facilitates consistency results given  $N = 1$  observation of a random graph with complex dependence. We have shown elsewhere [35] that when the neighborhood structure is known,  $M$ -estimators of canonical and curved exponential-family random graph models with local dependence and growing neighborhoods are consistent under weak conditions. While the neighborhood structure is observed in some applications [e.g., in multilevel networks, 24], it is unobserved in others. If the neighborhood structure is unobserved, the first and foremost question is whether it can be recovered with high probability given  $N = 1$  observation of a random graph with complex dependence. We present here the first consistency results which show that it is possible to do so as long as the data-generating exponential-family random graph model satisfies weak dependence and smoothness conditions. These consistency results are more challenging than consistency results in the special case of stochastic block models, because we cover exponential families with (a) countable support; (b) a wide range of dependencies within neighborhoods; and (c) a wide range of canonical and curved exponential-family parameterizations.

To recover the neighborhood structure along with the parameters given  $N = 1$  observation  $\mathbf{x}$  of  $\mathbf{X}$ , we consider the following restricted maximum likelihood estimator:

$$(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) \in \arg \max_{(\boldsymbol{\theta}, \mathbf{z}) \in \Theta_0 \times \mathbb{Z}_0} \ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{x})),$$

where

$$\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{x})) = \langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), s(\mathbf{x}) \rangle - \psi(\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}))$$

denotes the loglikelihood function of  $(\boldsymbol{\theta}, \mathbf{z}) \in \boldsymbol{\Theta}_0 \times \mathbb{Z}_0$  and  $\boldsymbol{\Theta}_0 \times \mathbb{Z}_0$  is a subset of  $\boldsymbol{\Theta} \times \mathbb{Z}$  to be specified. Computational implications are discussed in Section 6. We assume that the number of neighborhoods  $K$  is known and that both  $\boldsymbol{\theta}$  and  $\mathbf{z}$  are parameters, which is commonplace in the special case of stochastic block models [e.g., 3, 11, 1]. It is worth noting that the maximum likelihood estimator  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  is not unique, because the likelihood function is invariant to the labeling of neighborhoods. All following statements are therefore understood as statements about equivalence classes of neighborhood structures.

We call the maximum likelihood estimator  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  restricted, because we restrict maximum likelihood estimation to a subset  $\boldsymbol{\Theta}_0 \times \mathbb{Z}_0$  of  $\boldsymbol{\Theta} \times \mathbb{Z}$ . We need to do so, because without additional restrictions exponential families with local dependence can induce strong dependence and smoothness problems. To motivate the restrictions on  $\boldsymbol{\Theta} \times \mathbb{Z}$ , it is instructive to discuss the following concentration result, which is instrumental to deriving the main consistency results of the paper.

**Lemma 1.** *Suppose that a random graph is governed by an exponential family with local dependence and countable support  $\mathbb{X}$ . Let  $f : \mathbb{X} \mapsto \mathbb{R}$  be Lipschitz with respect to the Hamming metric  $d : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}_0^+$  defined by*

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i < j}^n \mathbb{1}_{x_{1,i,j} \neq x_{2,i,j}}, \quad (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{X} \times \mathbb{X},$$

*with Lipschitz coefficient  $\|f\|_{Lip} > 0$  and expectation  $\mathbb{E} f(\mathbf{X}) < \infty$ . Then there exists  $C > 0$  such that, for all  $n > 0$  and all  $t > 0$ ,*

$$\mathbb{P}(|f(\mathbf{X}) - \mathbb{E} f(\mathbf{X})| \geq t) \leq 2 \exp \left( -\frac{t^2}{C n^2 \|\mathcal{A}\|_\infty^4 \|f\|_{Lip}^2} \right),$$

*where  $\|\mathcal{A}\|_\infty = \max_{1 \leq k \leq K} |\mathcal{A}_k| > 0$  denotes the size of the largest data-generating neighborhood.*

The proof of Lemma 1 can be found in the appendix. The proof relies on concentration of measure inequalities for dependent random variables [21] and bounds mixing coefficients—which quantify the strength of dependence induced by exponential families with local dependence—in terms of  $\|\mathcal{A}\|_\infty$ .

Lemma 1 demonstrates that the probability mass of a function  $f(\mathbf{X})$  of a random graph concentrates around the corresponding expectation  $\mathbb{E} f(\mathbf{X})$  as long as the data-generating exponential family satisfies weak dependence and smoothness conditions. We are interested in applying Lemma 1 to concentrate exponential-family loglikelihood functions of the form  $\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{X})) = \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})}(\mathbf{X})$ . To make sure that the probability mass of  $\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})}(\mathbf{X})$  concentrates around the expectation  $\mathbb{E} \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})}(\mathbf{X})$ , we need to impose additional restrictions on  $\mathbb{Z}$  for at least two reasons. First of all, large neighborhoods can induce strong dependence, which weakens concentration results—as can be seen from the term  $\|\mathcal{A}\|_\infty$  in Lemma 1. Second, changes of edges in large neighborhoods can give rise to large changes of  $\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})}(\mathbf{x})$ , which weakens concentration results as well—as can be seen from the Lipschitz coefficient



$\|f\|_{\text{Lip}}$  in Lemma 1. Thus, to deal with strong dependence and smoothness problems, restrictions need to be imposed on the sizes of neighborhoods in  $\mathbb{Z}$ . An additional issue is that the unrestricted maximum likelihood estimator fails to exist with non-negligible probability [15, 28]. These observations motivate the following assumptions.

### 3.1 Assumptions

We assume that the data-generating natural parameter vector  $\boldsymbol{\eta}^* \in \Xi \subseteq \text{int}(\mathbb{N})$  is in the interior  $\text{int}(\mathbb{N})$  of the natural parameter space  $\mathbb{N}$ , which implies that the expectation  $\mathbb{E} s(\mathbf{X})$  exists [7, Theorem 2.2, pp. 34–35] and so does the expectation  $\mathbb{E} \ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{X}))$ , because

$$\mathbb{E} \ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{X})) = \langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), \mathbb{E} s(\mathbf{X}) \rangle - \psi(\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})) = \ell(\boldsymbol{\theta}, \mathbf{z}; \mathbb{E} s(\mathbf{X})).$$

Let  $\boldsymbol{\mu}(\boldsymbol{\eta}) = \mathbb{E}_{\boldsymbol{\eta}} s(\mathbf{X})$  be the mean-value parameter vector of an exponential family with natural parameter vector  $\boldsymbol{\eta} \equiv \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})$  and let  $\mathbb{M} = \text{rint}(\mathbb{C})$  be the mean-value parameter space, where  $\text{rint}(\mathbb{C})$  is the relative interior of the convex hull  $\mathbb{C} = \text{conv}\{s(\mathbf{x}) : \mathbf{x} \in \mathbb{X}\}$  of the set  $\{s(\mathbf{x}) : \mathbf{x} \in \mathbb{X}\}$ . It is well-known that in minimal exponential families the mapping between the relative interior of the mean-value and natural parameter space is one-to-one [7, Theorem 3.6, p. 74] and that all non-minimal exponential families can be reduced to minimal exponential families [7, Theorem 1.9, p. 13]. Denote by  $\boldsymbol{\mu}^* \equiv \boldsymbol{\mu}(\boldsymbol{\eta}^*)$  the data-generating mean-value parameter vector. For any  $\alpha > 0$ , let

$$\mathbb{M}(\alpha) = \{\boldsymbol{\mu} \in \mathbb{M} : |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}) - \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)| < \alpha |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|\}$$

be the subset of mean-value parameter vectors  $\boldsymbol{\mu} \in \mathbb{M}$  that are close to the data-generating mean-value parameter vector  $\boldsymbol{\mu}^* \in \mathbb{M}$  in the sense that  $|\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}) - \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)| < \alpha |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|$ . The advantage of introducing the subset  $\mathbb{M}(\alpha)$  of  $\mathbb{M}$  is that the main assumptions stated below can be weakened, because some of them need to hold on  $\mathbb{M}(\alpha)$ , but need not hold on  $\mathbb{M} \setminus \mathbb{M}(\alpha)$ .

The main assumptions can be stated as follows; note that conditions [C.2] and [C.3] are assumed to hold on  $\mathbb{M}(\alpha)$ , but need not hold on  $\mathbb{M} \setminus \mathbb{M}(\alpha)$ .

[C.1] For any fixed  $\mathbf{z} \in \mathbb{Z}$ , the map  $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \Xi$  is one-to-one and continuous on  $\boldsymbol{\Theta}$ .

[C.2] For any fixed  $\mathbf{z} \in \mathbb{Z}$  and any fixed  $\boldsymbol{\mu} \in \mathbb{M}(\alpha)$ , the loglikelihood function  $\ell(\boldsymbol{\theta}, \mathbf{z}; \boldsymbol{\mu})$  is upper semicontinuous on  $\boldsymbol{\Theta}$ .

[C.3] There exist  $A_1 > 0$  and  $n_1 > 0$  such that, for all  $n > n_1$ , all  $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \boldsymbol{\Theta} \times \boldsymbol{\Theta}$ , all  $\mathbf{z} \in \mathbb{Z}$ , and all  $\boldsymbol{\mu} \in \mathbb{M}(\alpha)$ ,

$$|\langle \boldsymbol{\eta}(\boldsymbol{\theta}_1, \mathbf{z}) - \boldsymbol{\eta}(\boldsymbol{\theta}_2, \mathbf{z}), \boldsymbol{\mu} \rangle| \leq A_1 \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|.$$

[C.4] There exist  $A_2 > 0$  and  $n_2 > 0$  such that, for all  $n > n_2$ , all  $(\boldsymbol{\theta}, \mathbf{z}) \in \boldsymbol{\Theta} \times \mathbb{Z}$ , and all  $(\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{X} \times \mathbb{X}$ ,

$$|\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), s(\mathbf{x}_1) - s(\mathbf{x}_2) \rangle| \leq A_2 d(\mathbf{x}_1, \mathbf{x}_2) L(\mathbf{z}),$$

where  $L(\mathbf{z})$  is the size of the largest neighborhood under  $\mathbf{z}$ .



[C.5] The data-generating parameters  $(\boldsymbol{\theta}^*, \mathbf{z}^*)$  are contained in  $\boldsymbol{\Theta}_0 \times \mathbb{Z}_0 \subseteq \boldsymbol{\Theta} \times \mathbb{Z}$ , where

- (a)  $\boldsymbol{\Theta}_0$  has dimension  $\dim(\boldsymbol{\theta}) \leq A n$  and can be covered by  $\exp(C n)$  closed balls  $\mathcal{B}(\boldsymbol{\theta}_l, B)$  with centers  $\boldsymbol{\theta}_l \in \boldsymbol{\Theta}$  and radius  $B > 0$ , i.e.,  $\boldsymbol{\Theta}_0 \subseteq \bigcup_{1 \leq l \leq \exp(C n)} \mathcal{B}(\boldsymbol{\theta}_l, B)$ , where  $A, B, C > 0$ .
- (b)  $\mathbb{Z}_0$  consists of all neighborhood structures for which the size of each of the  $K$  neighborhoods is bounded above by  $L$ , where  $K$  and  $L$  can increase as a function of the number of nodes  $n$ .

Corollaries 1 and 2 in Section 3.2 show that conditions [C.1]—[C.4] are satisfied by a wide range of canonical and curved exponential families with local dependence. Condition [C.1] along with the assumption that the exponential family is minimal ensures that  $\mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta}_1, \mathbf{z})} \neq \mathbb{P}_{\boldsymbol{\eta}(\boldsymbol{\theta}_2, \mathbf{z})}$  for all  $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$  given  $\mathbf{z} \in \mathbb{Z}$ ; note that all non-minimal exponential families can be reduced to minimal exponential families [7, Theorem 1.9, p. 13]. Conditions [C.2]—[C.4] are smoothness conditions. Condition [C.2] is a weak assumption: it is well-known that canonical exponential-family loglikelihood functions are upper semicontinuous [7, Lemma 5.3, p. 146] and it turns out that the most interesting curved exponential-family loglikelihood functions are upper semicontinuous as well, which is verified by Corollaries 1 and 2 in Section 3.2. Condition [C.3] imposes restrictions on how much  $\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})}(\mathbf{x})$  can change as a function of  $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})$ , whereas condition [C.4] imposes restrictions on how much  $\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})}(\mathbf{x})$  can change as a function of  $\mathbf{x}$ . Condition [C.3] is stated in terms of  $|\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|$  to accomodate both sparse and dense random graphs; we discuss the notion of sparse and dense random graphs in more detail in Remark 2 in Section 3.2. Condition [C.5](a) allows the dimension  $\dim(\boldsymbol{\theta})$  of the parameter space  $\boldsymbol{\Theta}_0$  to increase as a function of the number of nodes  $n$  and hence allows the model to be flexible while ensuring that  $\boldsymbol{\Theta}_0$  cannot be too large. We need these conditions, because we have  $N = 1$  observation and therefore cannot use conventional arguments to prove that estimators fall with high probability into compact subsets of the parameter space when the number of observations  $N$  is large [e.g., 2]. Condition [C.5](b) complements condition [C.4] and helps ensure that  $\log p_{\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})}(\mathbf{x})$  is not too sensitive to changes of  $\mathbf{x}$  by restricting the set of neighborhood structures to neighborhoods whose size is bounded above by  $L$ . The main consistency results of the paper, Proposition 1 and Theorem 1 in Section 3.2, impose restrictions on  $L$ .

## 3.2 Main consistency results

We discuss the main consistency results concerning the recovery of the neighborhood structure given  $N = 1$  observation of a random graph with complex dependence.

The recovery of the neighborhood structure is made possible by the following fundamental concentration result. The concentration result shows that with high probability the distribution parameterized by the restricted maximum likelihood estimator  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  is close to the distribution parameterized by the data-generating parameters  $(\boldsymbol{\theta}^*, \mathbf{z}^*)$  in terms of Kullback-Leibler divergence  $KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) = \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*) - \ell(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}; \boldsymbol{\mu}^*)$  provided that the

number of nodes  $n$  is sufficiently large. The result covers a wide range of canonical and curved exponential families with local dependence.

**Proposition 1.** *Suppose that  $N = 1$  observation of a random graph is generated by an exponential family with local dependence and countable support  $\mathbb{X}$  satisfying conditions [C.1]–[C.5]. Assume that, for all  $C_1 > 0$ , however large, there exists  $n_1 > 0$  such that, for all  $n > n_1$ ,*

$$|\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)| \geq C_1 n^{3/2} \|\mathcal{A}\|_\infty^2 L \sqrt{\log n}. \quad (4)$$

*Then there exist  $C > 0$ ,  $C_2 > 0$ , and  $n_2 > 0$  such that, for all  $n > n_2$ , with at least probability  $1 - 2 \exp(-\alpha^2 C_2 n \log n)$ , the restricted maximum likelihood estimator  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) \in \boldsymbol{\Theta}_0 \times \mathbb{Z}_0$  exists and, for all  $\epsilon > 0$ ,*

$$\mathbb{P}(KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) < \epsilon |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|) \geq 1 - 4 \exp(-\min(\alpha^2, \epsilon^2) C n \log n),$$

*where  $\alpha > 0$  is identical to the constant  $\alpha$  used in the construction of the subset  $\mathbb{M}(\alpha)$  of the mean-value parameter space  $\mathbb{M}$ .*

The concentration result in Proposition 1 paves the ground for the main consistency result. The consistency result is generic and covers a wide range of canonical and curved exponential families with local dependence. It states that the discrepancy between the estimated and data-generating neighborhood structure is small with high probability given  $N = 1$  observation of a random graph with complex dependence provided that the number of nodes  $n$  is sufficiently large. To define the discrepancy between the estimated and data-generating neighborhood structure, let  $\delta : \mathbb{Z} \times \mathbb{Z} \mapsto [0, n]$  be a discrepancy measure that is invariant to the labeling of neighborhoods. An example is given by  $\delta(\mathbf{z}^*, \hat{\mathbf{z}}) = \min_\pi \sum_{i=1}^n \mathbb{1}_{z_i^* \neq \pi(\hat{z}_i)}$ , the minimum Hamming distance between  $\mathbf{z}^*$  and  $\hat{\mathbf{z}}$ , where the minimum is taken with respect to all possible permutations  $\pi$  of  $\hat{\mathbf{z}}$ . The following consistency result holds for all discrepancy measures  $\delta : \mathbb{Z} \times \mathbb{Z} \mapsto [0, n]$  satisfying assumption (5) of the following result.

**Theorem 1.** *Suppose that  $N = 1$  observation of a random graph is generated by an exponential family with local dependence and countable support  $\mathbb{X}$  satisfying conditions [C.1]–[C.5]. If the random graph satisfies assumption (4) and there exist  $C_1 > 0$  and  $n_1 > 0$  such that, for all  $n > n_1$  and all  $(\boldsymbol{\theta}, \mathbf{z}) \in \boldsymbol{\Theta}_0 \times \mathbb{Z}_0$ ,*

$$KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\theta}, \mathbf{z}) \geq \frac{\delta(\mathbf{z}^*, \mathbf{z}) C_1 |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|}{n}, \quad (5)$$

*then there exist  $C > 0$ ,  $C_2 > 0$ , and  $n_2 > 0$  such that, for all  $n > n_2$ , with at least probability  $1 - 2 \exp(-\alpha^2 C_2 n \log n)$ , the restricted maximum likelihood estimator  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) \in \boldsymbol{\Theta}_0 \times \mathbb{Z}_0$  exists and, for all  $\epsilon > 0$ ,*

$$\mathbb{P}\left(\frac{\delta(\mathbf{z}^*, \hat{\mathbf{z}})}{n} < \epsilon\right) \geq 1 - 4 \exp(-\min(\alpha^2, \epsilon^2) C n \log n),$$

*where  $\alpha > 0$  is identical to the constant  $\alpha$  used in the construction of the subset  $\mathbb{M}(\alpha)$  of the mean-value parameter space  $\mathbb{M}$ .*

We discuss implications of Proposition 1 and Theorem 1, starting with a short comparison with stochastic block models (Remark 1) and then discussing assumption (4) (Remark 2) and its implications in terms of the sizes of neighborhoods (Remark 3) and the number of neighborhoods (Remark 4). We then proceed with a discussion of conditions [C.1]—[C.4] (Remark 5) and assumption (5) (Remark 6) and conclude with some comments on parameter estimation (Remark 7).

*Remark 1. Comparison with stochastic block models.* There is a large and growing body of consistency results on stochastic block models [e.g., 3, 11, 8, 30, 4, 1, 25, 31, 14]. In the language of stochastic block models, the consistency result in Theorem 1 is a weak consistency result in the sense that the discrepancy between the estimated and data-generating neighborhood structure is small with high probability. In contrast to stochastic block models, we cover exponential families with (a) countable support; (b) a wide range of dependencies within neighborhoods; and (c) a wide range of canonical and curved exponential-family parameterizations. These dependencies and parameterizations make theoretical results more challenging from a statistical point of view, but more relevant from a scientific point of view. However, these results come at a cost: in contrast to stochastic block models, we need to restrict the sizes of neighborhoods from above to deal with strong dependence and smoothness problems, as we pointed out in the discussion of Lemma 1. The restrictions on the sizes of neighborhoods are detailed in Remark 3.

*Remark 2. Assumption (4): sparse and dense random graphs.* Assumption (4) of Proposition 1 and Theorem 1 is stated in terms of the expected loglikelihood function  $|\mathbb{E} \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; s(\mathbf{X}))| = |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|$  to accomodate both sparse and dense random graphs. We call random graphs dense when  $|\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|$  grows as fast as  $n^2$  and sparse otherwise. It is worth noting that the conventional approach to quantifying the sparsity of random graphs is based on the expected number of edges—i.e., the expectation of the sufficient statistic of classic random graphs with independent and identically distributed edge variables—but in more general models it is desirable to quantify the sparsity of random graphs based on all sufficient statistics. Since it is not obvious how to combine all of the sufficient statistics, it is natural to use the expected loglikelihood function to quantify the sparsity of random graphs. Assumption (4) suggests that the random graph cannot be too sparse in the sense that the expected loglikelihood function  $|\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|$  cannot be too small. If, e.g.,  $\|\mathcal{A}\|_\infty$  and  $L$  grow as fast as  $(\log n)^{\gamma_1}$  ( $\gamma_1 > 0$ ) and  $(\log n)^{\gamma_2}$  ( $\gamma_2 > 0$ ), respectively, then  $|\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|$  must grow faster than  $n^{3/2} (\log n)^{2\gamma_1 + \gamma_2 + 1/2}$ . Therefore, the random graph must be denser than classic random graphs at the threshold of connectivity [6], where  $|\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|$  grows as fast as  $n \log n$ .

*Remark 3. Sizes of neighborhoods.* The sizes of neighborhoods in  $\mathbb{Z}_0$  cannot be too large, because changes of edges in large neighborhoods can give rise to large changes of  $\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{x})) = \log p_{\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})}(\mathbf{x})$ , which weakens concentration results, as we pointed out in the discussion of Lemma 1. In fact, assumption (4) implies that the size  $L$  of the largest possible neighborhood in  $\mathbb{Z}_0$  must satisfy

$$L \leq \frac{|\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|}{C_1 n^{3/2} \|\mathcal{A}\|_\infty^2 \sqrt{\log n}}.$$

Thus, in the best-case scenario when  $\|\mathcal{A}\|_\infty$  is small in the sense that  $\|\mathcal{A}\|_\infty$  grows at most as fast as  $(\log n)^\gamma$  ( $\gamma > 0$ ),  $L$  can grow at most as fast as  $n^{1/2}/(\log n)^{2\gamma+1/2}$ , assuming that the random graph is dense. In the worst-case scenario when  $\|\mathcal{A}\|_\infty$  grows as fast as  $L$ ,  $L$  can grow at most as fast as  $(n/\log n)^{1/6}$ .

*Remark 4. Number of neighborhoods.* The fact that the sizes of neighborhoods in  $\mathbb{Z}_0$  are bounded above by  $L$  implies that the number of neighborhoods  $K$  is bounded below by  $K \geq n/L$ . If, e.g.,  $L \leq n^{1/2}/(\log n)^{2\gamma+1/2}$  ( $\gamma > 0$ ), then  $K \geq n^{1/2}(\log n)^{2\gamma+1/2}$ . Compared with stochastic block models, the number of neighborhoods needs to grow at least as fast as in the high-dimensional stochastic block model of Choi et al. [11], where  $K$  can grow as fast as  $n^{1/2}$ , and may have to grow as fast as in the highest-dimensional stochastic block model of Rohe et al. [31], where  $K$  grows as fast as  $n$  (ignoring polylogarithmic terms).

*Remark 5. Conditions [C.1]–[C.4].* We show that conditions [C.1]–[C.4] are satisfied by a wide range of canonical and curved exponential families with local dependence. To ease the presentation, we consider dense random graphs, but the following results can be extended to sparse random graphs as long as the random graphs are not too sparse; see Remark 2.

We assume here that  $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \Xi$  is separable in the sense that  $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}) = \mathbf{A}(\mathbf{z}) \mathbf{b}(\boldsymbol{\theta})$ , where  $\mathbf{A} : \mathbb{Z} \mapsto \mathbb{R}^{\dim(\boldsymbol{\eta}) \times \dim(\mathbf{b})}$  and  $\mathbf{b} : \boldsymbol{\Theta} \mapsto \mathbb{R}^{\dim(\mathbf{b})}$ ; note that, e.g., the curved exponential-family parameterization described in Section 2.2 is separable, and so are many other canonical and curved exponential-family parameterizations. Since  $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \Xi$  is separable,  $\mathbf{A}(\mathbf{z})$  can be absorbed into the sufficient statistics vector, so that  $\boldsymbol{\eta} : \boldsymbol{\Theta} \mapsto \Xi$  can be considered as a function of  $\boldsymbol{\theta}$  and  $s : \mathbb{X} \times \mathbb{Z} \mapsto \mathbb{R}^{\dim(\boldsymbol{\eta})}$  can be considered as a function of  $\mathbf{x}$  and  $\mathbf{z}$ . As a result, we can write

$$\begin{aligned} \langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), \boldsymbol{\mu} \rangle &= \langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{\mu}(\mathbf{z}) \rangle = \sum_{k \leq l}^K \langle \boldsymbol{\eta}_{k,l}(\boldsymbol{\theta}), \boldsymbol{\mu}_{k,l}(\mathbf{z}) \rangle \\ \langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), s(\mathbf{x}) \rangle &= \langle \boldsymbol{\eta}(\boldsymbol{\theta}), s(\mathbf{x}, \mathbf{z}) \rangle = \sum_{k \leq l}^K \langle \boldsymbol{\eta}_{k,l}(\boldsymbol{\theta}), s_{k,l}(\mathbf{x}, \mathbf{z}) \rangle, \end{aligned}$$

where—in an abuse of notation—we write  $\boldsymbol{\mu}(\mathbf{z}) = \mathbf{A}(\mathbf{z})^\top \boldsymbol{\mu}$  ( $\boldsymbol{\mu} \in \mathbb{M}(\alpha)$ ) and  $s(\mathbf{x}, \mathbf{z}) = \mathbf{A}(\mathbf{z})^\top s(\mathbf{x})$  ( $s(\mathbf{x}) \in \mathbb{M}(\alpha)$ ). If, in addition,  $\mathbf{b}(\boldsymbol{\theta})$  is an affine function of  $\boldsymbol{\theta}$ , then  $\boldsymbol{\eta}(\boldsymbol{\theta})$  can be reduced to  $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$  and  $\boldsymbol{\eta}_{k,l}(\boldsymbol{\theta})$  can be reduced to  $\boldsymbol{\eta}_{k,l}(\boldsymbol{\theta}) = \boldsymbol{\theta}_{k,l}$  ( $k \leq l = 1, \dots, K$ ), in which case we call the exponential family canonical, otherwise we call the exponential family curved. In the following, we denote by  $L_k(\mathbf{z})$  the number of nodes in neighborhood  $k$  under neighborhood structure  $\mathbf{z} \in \mathbb{Z}_0$ .

The following result shows that conditions [C.1]–[C.4] are satisfied by all canonical exponential families with local dependence satisfying reasonable scaling and smoothness conditions.

**Corollary 1.** *Consider canonical exponential families with local dependence and countable support  $\mathbb{X}$ . Assume that  $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \Xi$  is separable with  $\dim(\boldsymbol{\theta}_{k,l}) < \infty$  ( $k \leq l = 1, \dots, K$ ) and that the random graph is dense. If there exist  $C_1 > 0$ ,  $C_2 > 0$ , and  $n_0 \geq 1$  such that, for all  $n > n_0$ ,*

[C.3<sup>★</sup>]  $\|\boldsymbol{\mu}_{k,l}(\mathbf{z})\|_\infty \leq C_1 L_k(\mathbf{z}) L_l(\mathbf{z})$  for all  $\mathbf{z} \in \mathbb{Z}_0$  and all  $\boldsymbol{\mu} \in \mathbb{M}(\alpha)$  ( $k \leq l = 1, \dots, K$ );

[C.4<sup>★</sup>]  $\sum_{k \leq l}^K \|s_{k,l}(\mathbf{x}_1, \mathbf{z}) - s_{k,l}(\mathbf{x}_2, \mathbf{z})\|_\infty \leq C_2 d(\mathbf{x}_1, \mathbf{x}_2) L(\mathbf{z})$  for all  $(\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{X} \times \mathbb{X}$  and all  $\mathbf{z} \in \mathbb{Z}_0$ ;

then conditions [C.1]—[C.4] are satisfied. If conditions [C.5] and (5) are satisfied as well, then the conclusions of Theorem 1 hold.

Condition [C.3<sup>★</sup>] is satisfied by all sufficient statistics that scale with the number of edge variables: e.g., the number of edges and transitive edges satisfy condition [C.3<sup>★</sup>] and so do all other sufficient statistics that count the number of pairs of nodes having certain properties or being related to other nodes in some specified form. Condition [C.4<sup>★</sup>] is satisfied by most sufficient statistics, including the number of edges and transitive edges.

We turn to curved exponential families with local dependence. We consider curved exponential families of densities of the form

$$p_{\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})}(\mathbf{x}) \propto \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), s(\mathbf{x}, \mathbf{z}) \rangle), \quad (6)$$

where

$$\langle \boldsymbol{\eta}(\boldsymbol{\theta}), s(\mathbf{x}, \mathbf{z}) \rangle = \sum_{k \leq l}^K \eta_{1,k,l}(\boldsymbol{\theta}) \sum_{i,j: z_i=k, z_j=l} x_{i,j} + \sum_{k=1}^K \sum_{t=1}^{T_k} \eta_{2,k,k,t}(\boldsymbol{\theta}) s_{k,k,t}(\mathbf{x}, \mathbf{z}),$$

where  $s_{k,k,t}(\mathbf{x}, \mathbf{z})$  are sufficient statistics that induce dependence within neighborhoods (e.g., in case  $\mathbb{X} = \{0, 1\}^{\binom{n}{2}}$ ,  $s_{k,k,t}(\mathbf{x}, \mathbf{z})$  may be the number of pairs of nodes with  $t$  edgewise shared partners in neighborhood  $k$ ). Here, the natural parameters are given by

$$\begin{aligned} \eta_{1,k,l}(\boldsymbol{\theta}) &= \theta_{1,k,l} \\ \eta_{2,k,k,t}(\boldsymbol{\theta}) &= \theta_{2,k} \left\{ \theta_{3,k} \left[ 1 - \left( 1 - \frac{1}{\theta_{3,k}} \right)^t \right] \right\}, \quad \theta_{3,k} > \frac{1}{2}, \quad T_k \geq 2. \end{aligned}$$

The following result shows that as long as the underlying geometric series converges, i.e., as long as  $\theta_{3,k} > 1/2$  ( $k = 1, \dots, K$ ), conditions [C.1]—[C.4] are satisfied. The result can be extended to other model terms, e.g., covariate terms.

**Corollary 2.** *Consider curved exponential families of the form (6) with local dependence and countable support  $\mathbb{X}$ . Assume that  $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \Xi$  is separable and that there exists  $B > 1/2$  such that  $1/2 < \theta_{3,k} < B$  ( $k = 1, \dots, K$ ) and that the random graph is dense. If there exist  $C_1 > 0$ ,  $C_2 > 0$ , and  $n_0 \geq 1$  such that, for all  $n > n_0$ ,*

[C.3<sup>★★</sup>]  $\sum_{t=1}^{T_k} |\mu_{k,k,t}(\mathbf{z})| \leq C_1 \binom{L_k(\mathbf{z})}{2}$  for all  $\mathbf{z} \in \mathbb{Z}_0$ , where  $\mu_{k,k,t}(\mathbf{z}) = \mathbb{E} s_{k,k,t}(\mathbf{X}, \mathbf{z})$ ;

[C.4<sup>★★</sup>]  $|\sum_{t=1}^{T_k} s_{k,k,t}(\mathbf{x}_1, \mathbf{z}) - \sum_{t=1}^{T_k} s_{k,k,t}(\mathbf{x}_2, \mathbf{z})| \leq C_2 d(\mathbf{x}_{1,k,k}, \mathbf{x}_{2,k,k}) L(\mathbf{z})$  for all  $(\mathbf{x}_{1,k,k}, \mathbf{x}_{2,k,k}) \in \mathbb{X}_{k,k}(\mathbf{z}) \times \mathbb{X}_{k,k}(\mathbf{z})$  and all  $\mathbf{z} \in \mathbb{Z}_0$ , where  $\mathbb{X}_{k,k}(\mathbf{z})$  denotes the set of all possible within-neighborhood subgraphs of neighborhood  $k$  under  $\mathbf{z} \in \mathbb{Z}_0$  ( $k = 1, \dots, K$ );

then conditions [C.1]–[C.4] are satisfied. If conditions [C.5] and (5) are satisfied as well, then the conclusions of Theorem 1 hold.

The most popular curved exponential families with geometrically weighted terms [38, 19, 18] satisfy conditions [C.3\*\*] and [C.4\*\*] of Corollary 2. Consider, e.g., geometrically weighted edgewise shared partner terms. In the case of geometrically weighted edgewise shared partner terms,  $T_k = L_k(\mathbf{z}) - 2$  and  $\sum_{t=1}^{T_k} s_{k,k,t}(\mathbf{x}, \mathbf{z})$  is the number of transitive edges in neighborhood  $k$ , hence conditions [C.3\*\*] and [C.4\*\*] are satisfied.

*Remark 6. Assumption (5).* Assumption (5) of Theorem 1 states that the Kullback-Leibler divergence of the distribution parameterized by  $(\boldsymbol{\theta}, \mathbf{z})$  from the distribution parameterized by  $(\boldsymbol{\theta}^*, \mathbf{z}^*)$  must increase with the discrepancy measure  $\delta(\mathbf{z}^*, \mathbf{z})$ . In the special case of stochastic block models, Choi et al. [11] and Rohe et al. [31] verified identifiability assumption (5) using the number of misclassified nodes as defined by Choi et al. [11] as a discrepancy measure, where the number of neighborhoods can grow as fast as  $n^{1/2}$  [11] and as fast as  $n$  (ignoring polylogarithmic terms) [31], respectively. In general, an application of the mean-value theorem to the expected loglikelihood function  $\ell(\boldsymbol{\eta}^*; \boldsymbol{\mu}^*) = \langle \boldsymbol{\eta}^*, \boldsymbol{\mu}^* \rangle - \psi(\boldsymbol{\eta}^*)$  shows that, for all  $\boldsymbol{\eta} \in \Xi \subseteq \text{int}(\mathbb{N})$ ,

$$KL(\boldsymbol{\eta}^*; \boldsymbol{\eta}) = \ell(\boldsymbol{\eta}^*; \boldsymbol{\mu}^*) - \ell(\boldsymbol{\eta}; \boldsymbol{\mu}^*) = \langle \boldsymbol{\eta}^* - \boldsymbol{\eta}, \boldsymbol{\mu}(\boldsymbol{\eta}^*) - \boldsymbol{\mu}(\dot{\boldsymbol{\eta}}) \rangle,$$

where  $\dot{\boldsymbol{\eta}} = \lambda \boldsymbol{\eta}^* + (1 - \lambda) \boldsymbol{\eta}$  ( $0 \leq \lambda \leq 1$ ); note that  $\dot{\boldsymbol{\eta}} \in \text{int}(\mathbb{N})$  since  $\boldsymbol{\eta}^* \in \text{int}(\mathbb{N})$  and  $\boldsymbol{\eta} \in \text{int}(\mathbb{N})$  and the natural parameter space  $\mathbb{N}$  is convex. Therefore, assumption (5) is satisfied as long as changes of neighborhoods give rise to large enough changes of mean-value and natural parameter vectors.

*Remark 7. Estimation of parameters.* The restricted maximum likelihood estimator estimates the parameter vector  $\boldsymbol{\theta}$  along with the neighborhood structure  $\mathbf{z}$ . We leave the study of the properties of estimators of  $\boldsymbol{\theta}$  to future research, but it is worth noting the following. If the neighborhoods are known [e.g., in multilevel networks, 24],  $M$ -estimators of canonical and curved exponential-family random graph models with local dependence and growing neighborhoods are consistent under weak conditions [35]. If the neighborhoods are unknown,  $M$ -estimators may not be consistent estimators of the data-generating parameters. Indeed, it is not too hard to see that, for any  $\mathbf{z} \neq \mathbf{z}^*$ —where  $\mathbf{z} \in \mathbb{Z}_0$  may be an estimate of  $\mathbf{z}^* \in \mathbb{Z}_0$ —the estimator

$$\hat{\boldsymbol{\theta}}(\mathbf{z}) = \arg \max_{\boldsymbol{\theta} \in \Theta_0} [\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{x})) - \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; s(\mathbf{x}))]$$

estimates

$$\dot{\boldsymbol{\theta}}(\mathbf{z}) = \arg \max_{\boldsymbol{\theta} \in \Theta_0} [\ell(\boldsymbol{\theta}, \mathbf{z}; \boldsymbol{\mu}^*) - \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)],$$

which is equivalent to minimizing the Kullback-Leibler divergence  $KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\theta}, \mathbf{z}) = \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*) - \ell(\boldsymbol{\theta}, \mathbf{z}; \boldsymbol{\mu}^*)$  with respect to  $\boldsymbol{\theta}$  given  $\mathbf{z} \in \mathbb{Z}_0$ . In other words,  $\hat{\boldsymbol{\theta}}(\mathbf{z})$  is an estimator of the parameter vector  $\dot{\boldsymbol{\theta}}(\mathbf{z})$  that is as close as possible to the data-generating parameter vector  $\boldsymbol{\theta}^*$  in terms of Kullback-Leibler divergence given  $\mathbf{z} \in \mathbb{Z}_0$ . These considerations suggest that  $\hat{\boldsymbol{\theta}}(\mathbf{z})$  may be a consistent estimator of  $\dot{\boldsymbol{\theta}}(\mathbf{z})$ , but in general  $\hat{\boldsymbol{\theta}}(\mathbf{z})$  is not a consistent estimator of  $\boldsymbol{\theta}^*$  unless  $\mathbf{z} = \mathbf{z}^*$  [35].



## 4 Simulation results

To demonstrate that the neighborhood structure can be recovered in practice, we simulate data from exponential families with neighborhood-dependent edge and transitive edge terms as described in Section 2.1. To estimate the model, note that the restricted maximum likelihood estimator is intractable, as discussed in Section 6. Here, we use Bayesian methods for small networks ( $n \leq 100$ ) and approximate maximum likelihood methods for large networks ( $n \gg 100$ ) in place of the intractable restricted maximum likelihood estimator.

### 4.1 Small networks

For small networks ( $n \leq 100$ ), Bayesian auxiliary-variable Markov chain Monte Carlo methods can be used to recover the neighborhood structure [33, 34]. We consider networks with  $n = 50$ ,  $n = 75$ , and  $n = 100$  nodes and  $K = 5$  neighborhoods  $\mathcal{A}_1, \dots, \mathcal{A}_K$  of equal size. The data-generating natural parameters are given by

$$\begin{aligned}\eta_{1,k,l} &= -\log\left(\frac{n - \min(\mathcal{A}_k, \mathcal{A}_l)}{3} - 1\right), \quad k < l = 1, \dots, K, \\ \eta_{1,k,k} &= -1, \quad \eta_{2,k,k} = 1, \quad k = 1, \dots, K,\end{aligned}$$

where the between-neighborhood natural parameters  $\eta_{1,k,l}$  have been chosen to ensure that, for each node, the expected number of edges between neighborhoods is 3. To deal with the so-called label-switching problem of Bayesian Markov chain Monte Carlo methods—which arises from the invariance of the likelihood function to the labeling of neighborhoods—we follow the Bayesian decision-theoretic approach of Stephens [39] and estimate neighborhood memberships by assigning each node to its maximum-posterior-probability neighborhood [33, 34].

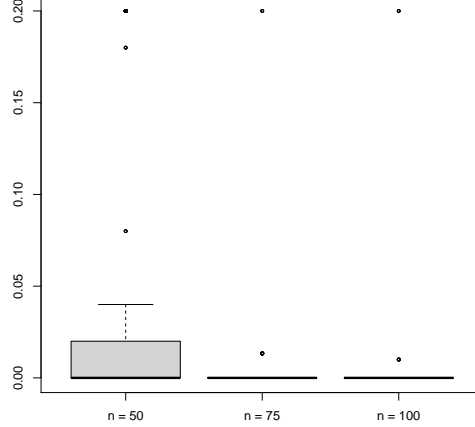
Figure 1 shows the fraction of misclassified nodes in terms of the normalized minimum Hamming distance  $\delta(\mathbf{z}^*, \hat{\mathbf{z}}) / n = \min_{\pi} \sum_{i=1}^n \mathbb{1}_{z_i^* \neq \pi(\hat{z}_i)} / n$  based on 100 simulated data sets with  $n = 50$ ,  $n = 75$ , and  $n = 100$  nodes and  $K = 5$  neighborhoods of equal size; we note that Bayesian methods are too time-consuming to be applied to more than 100 simulated data sets. Figure 1 suggests that the fraction of misclassified nodes is small in most data sets and decreases as the number of nodes increases from  $n = 50$  to  $n = 100$  and hence the sizes of the neighborhoods increases from 10 to 20.

### 4.2 Large networks

For large networks ( $n \gg 100$ ), Bayesian methods are too time-consuming and approximate methods have to be used. We demonstrate them here and elaborate on them elsewhere. The approximate methods are based on the idea that as long as the neighborhoods are not too large and the random graph is not too sparse, the  $\binom{K}{2}$  between-neighborhood subgraphs dominate the random graph. Therefore, despite the fact that exponential families induce dependence within neighborhoods and hence have an added value relative to stochastic block models—as pointed out in Section 2.4—most of the random graph is governed by the



Figure 1: Fraction of misclassified nodes based on 100 simulated data sets with  $n = 50$ ,  $n = 75$ , and  $n = 100$  nodes and  $K = 5$  neighborhoods of equal size, where the model is estimated by Bayesian methods.



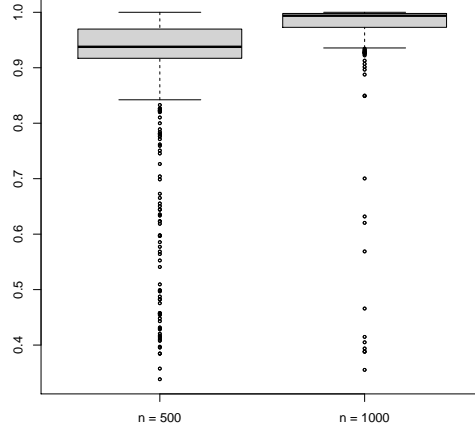
same probability law as random graphs governed by stochastic block models. Thus, one can estimate the neighborhood structure by using approximate methods based on stochastic block models. Stochastic block models admit the estimation of neighborhood structure from large networks [e.g., 30, 5, 40]. To demonstrate, we consider approximate methods based on the following two-step estimation approach. In the first step, we estimate the neighborhood structure by assuming that  $\eta_{2,k,k} = 0$  ( $k = 1, \dots, K$ )—in which case the exponential family with local dependence reduces to stochastic block models—and estimating the neighborhood structure by using variational methods for stochastic block models described in Vu et al. [40]. In the second step, we estimate parameters under the assumption that  $\eta_{2,k,k} \neq 0$  ( $k = 1, \dots, K$ ) conditional on the estimated neighborhood structure by using Monte Carlo maximum likelihood methods described by Hunter and Handcock [19].

We consider networks with  $K = 50$  neighborhoods of equal size, where the size is either 10 or 20, so that  $n = 500$  or  $n = 1,000$ . The data-generating natural parameters are given by

$$\begin{aligned} \eta_{1,k,l} &= -\log \left( \frac{n - \min(\mathcal{A}_k, \mathcal{A}_l)}{2.5} - 1 \right), \quad k < l = 1, \dots, K, \\ \eta_{1,k,k} &= -\log \left( \frac{n - \min(\mathcal{A}_k, \mathcal{A}_l)}{5} - 1 \right), \quad k = 1, \dots, K, \\ \eta_{2,k,k} &= 1, \quad k = 1, \dots, K. \end{aligned}$$

Figure 2 shows the agreement of the estimated and data-generating neighborhood structure in terms of Yule’s  $\phi$ -coefficient [10] based on 1,000 simulated data sets with  $n = 500$  and  $n = 1,000$  nodes with  $K = 50$  neighborhoods of equal size; note that the minimum normalized Hamming distance cannot be computed when  $K \gg 5$ , because the minimization over all possible  $K!$  permutations is infeasible when  $K \gg 5$ . Figure 2 demonstrates that

Figure 2: Agreement of estimated and data-generating neighborhood structure in terms of Yule’s  $\phi$ -coefficient based on 1,000 simulated data sets with  $n = 500$  and  $n = 1,000$  nodes with  $K = 50$  neighborhoods of equal size, where the model is estimated by approximate maximum likelihood methods.



the agreement is high in most data sets and increases as the number of nodes increases from  $n = 500$  to  $n = 1,000$  and hence the sizes of the neighborhoods increases from 10 to 20.

## 5 Proofs of main consistency results

We prove the main consistency results, Proposition 1 and Theorem 1. To prove them, we need two additional lemmas, Lemmas 2 and 3. The proofs of Lemmas 1, 2, and 3 are delegated to the appendix along with the proofs of Corollaries 1 and 2.

To state Lemmas 2 and 3, note that the data-generating natural parameter vector  $\eta^* \in \Xi \subseteq \text{int}(\mathbb{N})$  is in the interior  $\text{int}(\mathbb{N})$  of the natural parameter space  $\mathbb{N}$ . Therefore, the expectation  $\mathbb{E} s(\mathbf{X})$  exists [7, Theorem 2.2, pp. 34–35] and so does the expectation  $\mathbb{E} \ell(\theta, \mathbf{z}; s(\mathbf{X})) = \ell(\theta, \mathbf{z}; \mathbb{E} s(\mathbf{X}))$ . Let

$$\mathbb{X}(\alpha) = \{\mathbf{x} \in \mathbb{X} : |\ell(\theta^*, \mathbf{z}^*; s(\mathbf{x})) - \ell(\theta^*, \mathbf{z}^*; \mu^*)| < \alpha |\ell(\theta^*, \mathbf{z}^*; \mu^*)|\}$$

be the subset of  $\mathbf{x} \in \mathbb{X}$  such that  $s(\mathbf{x}) \in \mathbb{M}(\alpha)$ , where  $\alpha > 0$  is identical to the constant  $\alpha$  used in the construction of the subset  $\mathbb{M}(\alpha)$  of the mean-value parameter space  $\mathbb{M}$ .

Lemma 2 shows that the event  $\mathbf{X} \in \mathbb{X}(\alpha)$  occurs with high probability provided that the number of nodes  $n$  is sufficiently large and hence all probability statements in Proposition 1 and Theorem 1 can be restricted to the high-probability subset  $\mathbb{X}(\alpha)$  of  $\mathbb{X}$ .

**Lemma 2.** *Suppose that  $N = 1$  observation of a random graph is generated by an exponential family with local dependence and countable support  $\mathbb{X}$  satisfying condition [C.4] along with assumption (4). Then there exist  $C > 0$  and  $n_0 > 0$  such that, for all  $n > n_0$ ,*

$$\mathbb{P}(\mathbf{X} \in \mathbb{X}(\alpha)) \geq 1 - 2 \exp(-\alpha^2 C n \log n),$$

where  $\alpha > 0$  is identical to the constant  $\alpha$  used in the construction of the subset  $\mathbb{M}(\alpha)$  of the mean-value parameter space  $\mathbb{M}$ .

Lemma 3 shows that in the event  $\mathbf{X} \in \mathbb{X}(\alpha)$ , the restricted maximum likelihood estimator  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  exists, which implies that the restricted maximum likelihood estimator  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  exists with high probability provided that the number of nodes  $n$  is sufficiently large by Lemma 2.

**Lemma 3.** *Suppose that  $N = 1$  observation of a random graph is generated by an exponential family with local dependence and countable support  $\mathbb{X}$  satisfying conditions [C.2] and [C.4] along with assumption (4). Then the following statements hold:*

- (a) *For all  $\mathbf{x} \in \mathbb{X}(\alpha)$ , the restricted maximum likelihood estimator  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  exists;*
- (b) *There exist  $C > 0$  and  $n_0 > 0$  such that, for all  $n > n_0$ , the restricted maximum likelihood estimator  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  exists with at least probability  $1 - 2 \exp(-\alpha^2 C n \log n)$ ;*

where  $\alpha > 0$  is identical to the constant  $\alpha$  used in the construction of the subset  $\mathbb{M}(\alpha)$  of the mean-value parameter space  $\mathbb{M}$ .

Armed with Lemmas 2 and 3, we can prove Proposition 1 and Theorem 1.

PROOF OF PROPOSITION 1. Throughout, to ease the presentation, we use the short-hand expression

$$u(n) = |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|.$$

By Lemma 2, there exist  $C_0 > 0$  and  $n_0 > 0$  such that, for all  $n > n_0$ ,

$$\mathbb{P}(\mathbb{X} \setminus \mathbb{X}(\alpha)) \leq 2 \exp(-\alpha^2 C_0 n \log n).$$

Thus, all following arguments can be restricted to the high-probability subset  $\mathbb{X}(\alpha)$  of  $\mathbb{X}$ . It is therefore convenient to bound the probability of the event  $KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) \geq \epsilon u(n)$  by using a divide- and conquer strategy based on the inequality

$$\begin{aligned} & \mathbb{P}\left(KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) \geq \epsilon u(n)\right) \\ & \leq \mathbb{P}\left(KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) \geq \epsilon u(n) \cap \mathbb{X}(\alpha)\right) + \mathbb{P}(\mathbb{X} \setminus \mathbb{X}(\alpha)). \end{aligned} \tag{7}$$

The advantage of doing so is that we can confine attention to observations  $s(\mathbf{x}) \in \mathbb{M}(\alpha)$  that fall into well-behaved subsets  $\mathbb{M}(\alpha)$  of the mean-value parameter space  $\mathbb{M}$  satisfying conditions [C.2] and [C.3]. Observe that conditions [C.2] and [C.3] are assumed to hold on  $\mathbb{M}(\alpha)$ , but need not hold on  $\mathbb{M} \setminus \mathbb{M}(\alpha)$ .

To bound the probability of the event  $KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) \geq \epsilon u(n) \cap \mathbb{X}(\alpha)$ , note that, for any  $\mathbf{x} \in \mathbb{X}(\alpha)$ , the restricted maximum likelihood estimator  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  exists by Lemma 3 and that

$$KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) = \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*) - \ell(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}; \boldsymbol{\mu}^*) \geq 0.$$

Since  $(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{z}}) \in \boldsymbol{\Theta}_0 \times \mathbb{Z}_0$  maximizes  $\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{x}))$  and  $(\boldsymbol{\theta}^*, \mathbf{z}^*) \in \boldsymbol{\Theta}_0 \times \mathbb{Z}_0$ , we have

$$\begin{aligned} & \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*) + [\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; s(\mathbf{x})) - \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)] \\ & \leq \ell(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{z}}; \boldsymbol{\mu}^*) + [\ell(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{z}}; s(\mathbf{x})) - \ell(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{z}}; \boldsymbol{\mu}^*)] \end{aligned}$$

and hence  $KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{z}})$  can be bounded above as follows:

$$\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*) - \ell(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{z}}; \boldsymbol{\mu}^*) \leq 2 \max_{\mathbf{z} \in \mathbb{Z}_0} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} |\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{x})) - \ell(\boldsymbol{\theta}, \mathbf{z}; \boldsymbol{\mu}^*)|.$$

Choose any  $\rho > 0$  satisfying  $0 < \rho < \epsilon / (12 A_1)$ , where  $A_1 > 0$  is equal to the constant  $A_1 > 0$  in condition [C.3]. By condition [C.5], there exist  $A, B, C > 0$  such that the  $\dim(\boldsymbol{\theta}) \leq A n$ -dimensional parameter space  $\boldsymbol{\Theta}_0 \subseteq \boldsymbol{\Theta}$  can be covered by  $\exp(C n)$  closed balls with centers  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$  and radius  $B > 0$ . Each of the  $\exp(C n)$  balls with radius  $B > 0$  can be covered by

$$\left( \frac{4B + \rho}{\rho} \right)^{\dim(\boldsymbol{\theta})}$$

balls  $\mathcal{B}(\boldsymbol{\theta}, \rho)$  with centers  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$  and radius  $\rho > 0$ . Therefore,  $\boldsymbol{\Theta}_0 \subseteq \bigcup_{1 \leq l \leq L} \mathcal{B}(\boldsymbol{\theta}_l, \rho)$  can be covered by  $L$  balls  $\mathcal{B}(\boldsymbol{\theta}_l, \rho)$  with centers  $\boldsymbol{\theta}_l \in \boldsymbol{\Theta}$  and radius  $\rho > 0$ , where  $L$  is bounded above by

$$L \leq \exp \left( A \log \left( \frac{4B + \rho}{\rho} \right) n + C n \right). \quad (8)$$

As a result, we can write

$$\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*) - \ell(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{z}}; \boldsymbol{\mu}^*) \leq 2 \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} \sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_l, \rho)} |\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{x})) - \ell(\boldsymbol{\theta}, \mathbf{z}; \boldsymbol{\mu}^*)|.$$

Collecting terms shows that

$$\begin{aligned} & \mathbb{P} \left( KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{z}}) \geq \epsilon u(n) \cap \mathbb{X}(\alpha) \right) \\ & = \mathbb{P} \left( \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*) - \ell(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{z}}; \boldsymbol{\mu}^*) \geq \epsilon u(n) \cap \mathbb{X}(\alpha) \right) \\ & \leq \mathbb{P} \left( \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} \sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_l, \rho)} |\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{X})) - \ell(\boldsymbol{\theta}, \mathbf{z}; \boldsymbol{\mu}^*)| \geq \frac{\epsilon u(n)}{2} \cap \mathbb{X}(\alpha) \right). \end{aligned}$$

To bound the probability of the max-sup of deviations of the form  $|\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{X})) - \ell(\boldsymbol{\theta}, \mathbf{z}; \boldsymbol{\mu}^*)|$ , observe that, for any  $\mathbf{x} \in \mathbb{X}(\alpha)$ , the deviation reduces to

$$|\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{x})) - \ell(\boldsymbol{\theta}, \mathbf{z}; \boldsymbol{\mu}^*)| = |\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), s(\mathbf{x}) \rangle - \langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), \boldsymbol{\mu}^* \rangle|,$$

because  $\psi(\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}))$  cancels. Consider any  $\mathbf{z} \in \mathbb{Z}_0$  and any of the  $L$  balls  $\mathcal{B}(\boldsymbol{\theta}_l, \rho)$  that make up the cover  $\bigcup_{1 \leq l \leq L} \mathcal{B}(\boldsymbol{\theta}_l, \rho)$  of  $\boldsymbol{\Theta}_0$ . Let

$$\dot{\boldsymbol{\theta}}_l(\mathbf{z}) = \arg \max_{\boldsymbol{\theta} \in \text{cl } \mathcal{B}(\boldsymbol{\theta}_l, \rho)} \ell(\boldsymbol{\theta}, \mathbf{z}; \boldsymbol{\mu}^*),$$

where the subscript  $l$  is added to indicate the closed ball  $\text{cl } \mathcal{B}(\boldsymbol{\theta}_l, \rho)$  that contains  $\dot{\boldsymbol{\theta}}_l(\mathbf{z})$ . Observe that, for any  $\mathbf{z} \in \mathbb{Z}_0$ ,  $\ell(\boldsymbol{\theta}, \mathbf{z}; \boldsymbol{\mu}^*)$  is upper semicontinuous on  $\text{cl } \mathcal{B}(\boldsymbol{\theta}_l, \rho)$  by condition [C.2] and hence assumes a maximum on  $\text{cl } \mathcal{B}(\boldsymbol{\theta}_l, \rho)$ . Thus, for any  $\mathbf{z} \in \mathbb{Z}_0$ , the maximizer  $\dot{\boldsymbol{\theta}}_l(\mathbf{z})$  exists and is unique by condition [C.1] and the assumption that the exponential family is minimal, which can be assumed without loss [7, Theorem 1.9, p. 13]. The triangle inequality shows that, for any  $\mathbf{x} \in \mathbb{X}(\alpha)$ , any  $\mathbf{z} \in \mathbb{Z}_0$ , any  $\boldsymbol{\theta} \in \text{cl } \mathcal{B}(\boldsymbol{\theta}_l, \rho)$ , and any  $\dot{\boldsymbol{\theta}}_l(\mathbf{z}) \in \text{cl } \mathcal{B}(\boldsymbol{\theta}_l, \rho)$ ,

$$\begin{aligned} |\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{x})) - \ell(\boldsymbol{\theta}, \mathbf{z}; \boldsymbol{\mu}^*)| &= |\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), s(\mathbf{x}) \rangle - \langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), \boldsymbol{\mu}^* \rangle| \\ &\leq |\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), s(\mathbf{x}) \rangle - \langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{x}) \rangle| \\ &\quad + |\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{x}) \rangle - \langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), \boldsymbol{\mu}^* \rangle| \\ &\quad + |\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), \boldsymbol{\mu}^* \rangle - \langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), \boldsymbol{\mu}^* \rangle|. \end{aligned}$$

A union bound over the three terms on the right-hand side of the inequality above shows that

$$\begin{aligned} &\mathbb{P} \left( \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} \sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_l, \rho)} |\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{X})) - \ell(\boldsymbol{\theta}, \mathbf{z}; \boldsymbol{\mu}^*)| \geq \frac{\epsilon u(n)}{2} \cap \mathbb{X}(\alpha) \right) \\ &\leq \mathbb{P} \left( \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} \sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_l, \rho)} |\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}) - \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{X}) \rangle| \geq \frac{\epsilon u(n)}{6} \cap \mathbb{X}(\alpha) \right) \\ &\quad + \mathbb{P} \left( \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} \sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_l, \rho)} |\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{X}) - \boldsymbol{\mu}^* \rangle| \geq \frac{\epsilon u(n)}{6} \cap \mathbb{X}(\alpha) \right) \\ &\quad + \mathbb{P} \left( \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} \sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_l, \rho)} |\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}) - \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), \boldsymbol{\mu}^* \rangle| \geq \frac{\epsilon u(n)}{6} \cap \mathbb{X}(\alpha) \right). \end{aligned}$$

We bound the last three terms on the right-hand side of the inequality above one by one.

**First term.** The first term can be bounded by using condition [C.3], which implies that there exist  $A_1 > 0$  and  $n_1 > 0$  such that, for any  $n > n_1$ , any  $\mathbf{x} \in \mathbb{X}(\alpha)$ , any  $\mathbf{z} \in \mathbb{Z}_0$ , any  $\boldsymbol{\theta} \in \text{cl } \mathcal{B}(\boldsymbol{\theta}_l, \rho)$ , and any  $\dot{\boldsymbol{\theta}}_l(\mathbf{z}) \in \text{cl } \mathcal{B}(\boldsymbol{\theta}_l, \rho)$ ,

$$|\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}) - \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{x}) \rangle| \leq A_1 \|\boldsymbol{\theta} - \dot{\boldsymbol{\theta}}_l(\mathbf{z})\|_2 u(n).$$

Since both  $\boldsymbol{\theta}$  and  $\dot{\boldsymbol{\theta}}_l(\mathbf{z})$  are contained in the ball  $\text{cl } \mathcal{B}(\boldsymbol{\theta}_l, \rho)$ , an application of the triangle inequality shows that

$$A_1 \|\boldsymbol{\theta} - \dot{\boldsymbol{\theta}}_l(\mathbf{z})\|_2 u(n) \leq A_1 2 \rho u(n) < \frac{\epsilon u(n)}{6},$$

where we used the fact that  $\rho > 0$  satisfies  $0 < \rho < \epsilon / (12 A_1)$  by construction. As a result, for all  $n > n_1$ , we have

$$\mathbb{P} \left( \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} \sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_l, \rho)} |\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}) - \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{X}) \rangle| \geq \frac{\epsilon u(n)}{6} \cap \mathbb{X}(\alpha) \right) = 0.$$

**Second term.** We are interested in bounding the probability of deviations of the form  $|\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{X}) - \boldsymbol{\mu}^* \rangle|$ . We make two observations. First, observe that, for any  $\mathbf{x} \in \mathbb{X}(\alpha)$ ,

$$\begin{aligned} & \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} \sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_l, \rho)} |\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{x}) - \boldsymbol{\mu}^* \rangle| \\ &= \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} |\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{x}) - \boldsymbol{\mu}^* \rangle|, \end{aligned}$$

which implies that

$$\begin{aligned} & \mathbb{P} \left( \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} \sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_l, \rho)} |\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{X}) - \boldsymbol{\mu}^* \rangle| \geq \frac{\epsilon u(n)}{6} \cap \mathbb{X}(\alpha) \right) \\ &= \mathbb{P} \left( \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} |\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{X}) - \boldsymbol{\mu}^* \rangle| \geq \frac{\epsilon u(n)}{6} \cap \mathbb{X}(\alpha) \right). \end{aligned}$$

Second, bounding the probability of deviations of the form  $|\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{X}) - \boldsymbol{\mu}^* \rangle|$  is equivalent to bounding the probability of deviations of the form  $|f(\mathbf{X}) - \mathbb{E} f(\mathbf{X})|$ , where

$$f(\mathbf{X}) = \langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{X}) \rangle, \quad \mathbb{E} f(\mathbf{X}) = \langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), \boldsymbol{\mu}^* \rangle.$$

Here,  $f : \mathbb{X} \mapsto \mathbb{R}$  is considered as a function of  $\mathbf{X}$  for fixed  $(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}) \in \boldsymbol{\Theta}_0 \times \mathbb{Z}_0$ . To bound the probability of deviations of the form  $|f(\mathbf{X}) - \mathbb{E} f(\mathbf{X})|$ , observe that by condition [C.4] there exist  $A_2 > 0$  and  $n_2 > 0$  such that, for all  $n > n_2$ , the Lipschitz coefficient of  $f(\mathbf{X})$  satisfies  $\|f\|_{\text{Lip}} \leq A_2 L$ . Thus, by applying Lemma 1 to deviations of size  $t = \epsilon u(n) / 6$  along with a union bound over the  $|\mathbb{Z}_0|$  neighborhood structures and all  $L$  balls that make up the cover  $\bigcup_{1 \leq l \leq L} \mathcal{B}(\boldsymbol{\theta}_l, \rho)$  of  $\boldsymbol{\Theta}_0$ , there exists  $C_1 > 0$  such that, for all  $\epsilon > 0$  and all  $n > n_2$ ,

$$\begin{aligned} & \mathbb{P} \left( \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} |\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{X}) - \boldsymbol{\mu}^* \rangle| \geq \frac{\epsilon u(n)}{6} \cap \mathbb{X}(\alpha) \right) \\ &\leq \mathbb{P} \left( \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} |\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{X}) - \boldsymbol{\mu}^* \rangle| \geq \frac{\epsilon u(n)}{6} \right) \\ &\leq 2 \exp \left( -\frac{\epsilon^2 u(n)^2}{36 C_1 n^2 \|\mathcal{A}\|_\infty^4 L^2} + \log |\mathbb{Z}_0| + \log L \right). \end{aligned}$$

To bound the exponential term, observe that by assumption (4) of Proposition 1 there exists, for all  $M > 0$ , however large,  $n_3 > 0$  such that, for all  $n > n_3$ ,

$$u(n) \geq M n^{3/2} \|\mathcal{A}\|_\infty^2 L \sqrt{\log n}.$$

Therefore, for all  $n > n_3$ , the three terms in the exponent are bounded above by

$$\begin{aligned} & -\frac{\epsilon^2 u(n)^2}{36 C_1 n^2 \|\mathcal{A}\|_\infty^4 L^2} + \log |\mathbb{Z}_0| + \log L \\ &\leq -\frac{\epsilon^2 u(n)^2}{36 C_1 n^2 \|\mathcal{A}\|_\infty^4 L^2} + \left[ 1 + A \log \left( \frac{4B + \rho}{\rho} \right) + C \right] n \log n, \end{aligned}$$

where we used  $\log |\mathbb{Z}_0| \leq n \log K$  and  $\log L \leq (A \log(4B + \rho)/\rho + C)n$  by (8). Since  $M > 0$  can be chosen as large as desired, we can choose

$$M > \sqrt{36 C_1 C_2 \left[ 1 + A \log \left( \frac{4B + \rho}{\rho} \right) + C \right]},$$

where  $C_2 > 0$  is chosen so that  $C_2 \epsilon^2 > 1$ . Hence there exists  $C_3 > 0$  such that, for all  $n > n_3$ ,

$$-\frac{\epsilon^2 u(n)^2}{36 C_1 n^2 \|\mathcal{A}\|_\infty^4 L^2} + \left[ 1 + A \log \left( \frac{4B + \rho}{\rho} \right) + C \right] n \log n \leq -\epsilon^2 C_3 n \log n.$$

Collecting terms shows that, for all  $n > n_3$ ,

$$\begin{aligned} \mathbb{P} \left( \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} \sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_l, \rho)} |\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}), s(\mathbf{X}) - \boldsymbol{\mu}^* \rangle| \geq \frac{\epsilon u(n)}{6} \cap \mathbb{X}(\alpha) \right) \\ \leq 2 \exp(-\epsilon^2 C_3 n \log n). \end{aligned}$$

**Third term.** The third term can be bounded along the same lines as the first term, which implies that there exists  $n_4 > 0$  such that, for all  $n > n_4$ ,

$$\mathbb{P} \left( \max_{\mathbf{z} \in \mathbb{Z}_0} \max_{1 \leq l \leq L} \sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_l, \rho)} |\langle \boldsymbol{\eta}(\dot{\boldsymbol{\theta}}_l(\mathbf{z}), \mathbf{z}) - \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), \boldsymbol{\mu}^* \rangle| \geq \frac{\epsilon u(n)}{6} \cap \mathbb{X}(\alpha) \right) = 0.$$

**Conclusion.** Using (7) and collecting terms shows that there exists  $C > 0$  such that, for all  $\epsilon > 0$  and all  $n > \max(n_0, n_1, n_2, n_3, n_4)$ ,

$$\begin{aligned} \mathbb{P} \left( KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) \geq \epsilon u(n) \right) &\leq 2 \exp(-\alpha^2 C_0 n \log n) \\ &+ 2 \exp(-\epsilon^2 C_3 n \log n) \leq 4 \exp(-\min(\alpha^2, \epsilon^2) C n \log n). \end{aligned}$$

**PROOF OF THEOREM 1.** By assumption (5) of Theorem 1, there exist  $C_1 > 0$  and  $n_1 > 0$  such that, for all  $n > n_1$ ,

$$KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) \geq \frac{\delta(\mathbf{z}^*, \hat{\mathbf{z}}) C_1 |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|}{n}$$

provided  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  exists. By Proposition 1, there exist  $C_2 > 0$  and  $n_2 > 0$  such that, for all  $\epsilon > 0$  and all  $n > n_2$ , the event

$$KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) < \epsilon C_1 |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|$$

occurs with at least probability

$$1 - 4 \exp(-\min(\alpha^2, \epsilon^2) C_2 n \log n). \quad (9)$$

Therefore, for all  $\epsilon > 0$  and all  $n > \max(n_1, n_2)$ , with at least probability (9), we observe the event

$$\frac{\delta(\mathbf{z}^*, \hat{\mathbf{z}}) C_1 |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|}{n} \leq KL(\boldsymbol{\theta}^*, \mathbf{z}^*; \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}) < \epsilon C_1 |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|,$$

i.e., the event  $\delta(\mathbf{z}^*, \hat{\mathbf{z}})/n < \epsilon$ .



## 6 Discussion

Here, and elsewhere [35], we have taken first steps to demonstrate that—while statistical inference for exponential-family random graph models without additional structure is problematic [15, 32, 9, 36]—statistical inference for exponential-family random graph models with additional structure makes sense. It goes without saying that numerous open problems remain, ranging from probabilistic problems (e.g., understanding properties of probability models) and statistical problems (e.g., understanding properties of statistical methods) to computational problems (e.g., the development of computational methods for large networks).

One important problem is that the maximum likelihood estimator discussed here is at least as intractable as maximum likelihood estimators in the special case of stochastic block models [11, 31]. The intractability stems in part from the fact that the neighborhood structure is unknown and the number of possible neighborhood structures is large and in part from the fact that the likelihood function is intractable even when the neighborhood structure is known owing to local dependence. There do exist Bayesian auxiliary-variable methods for small networks [33, 34] and promising directions for methods for large networks, as pointed out in Section 4. Future work will focus on developing computational methods for large networks.

## Acknowledgements

The author acknowledges support from the National Science Foundation (NSF award DMS-1513644).

## References

- [1] Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013), “Pseudo-likelihood methods for community detection in large sparse networks,” *The Annals of Statistics*, 41, 2097–2122.
- [2] Berk, R. H. (1972), “Consistency and asymptotic normality of MLE’s for exponential models,” *The Annals of Mathematical Statistics*, 43, 193–204.
- [3] Bickel, P. J., and Chen, A. (2009), “A nonparametric view of network models and Newman-Girvan and other modularities,” in *Proceedings of the National Academy of Sciences*, Vol. 106, pp. 21068–21073.
- [4] Bickel, P. J., Chen, A., and Levina, E. (2011), “The method of moments and degree distributions for network models,” *The Annals of Statistics*, 39, 2280–2301.
- [5] Bickel, P. J., Choi, D., Chang, X., and Zhang, H. (2013), “Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels,” *The Annals of Statistics*, 41, 1922–1943.

- [6] Bollobás, B. (1998), *Modern Graph Theory*, New York: Springer-Verlag.
- [7] Brown, L. (1986), *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*, Hayworth, CA, USA: Institute of Mathematical Statistics.
- [8] Celisse, A., Daudin, J. J., and Pierre, L. (2012), “Consistency of maximum-likelihood and variational estimators in the stochastic block model,” *Electronic Journal of Statistics*, 6, 1847–1899.
- [9] Chatterjee, S., and Diaconis, P. (2013), “Estimating and understanding exponential random graph models,” *The Annals of Statistics*, 41, 2428–2461.
- [10] Chedzoy, O. B. (2004), “Phi-Coefficient,” in *Encyclopedia of Statistical Sciences*, Wiley.
- [11] Choi, D. S., Wolfe, P. J., and Airolidi, E. M. (2012), “Stochastic blockmodels with growing number of classes,” *Biometrika*, 99, 273–284.
- [12] Diaconis, P., Chatterjee, S., and Sly, A. (2011), “Random graphs with a given degree sequence,” *The Annals of Applied Probability*, 21, 1400–1435.
- [13] Frank, O., and Strauss, D. (1986), “Markov graphs,” *Journal of the American Statistical Association*, 81, 832–842.
- [14] Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2016), “Achieving Optimal Misclassification Proportion in Stochastic Block Model,” *The Annals of Statistics*, in press.
- [15] Handcock, M. (2003), “Assessing degeneracy in statistical models of social networks,” Tech. rep., Center for Statistics and the Social Sciences, University of Washington, <http://www.csss.washington.edu/Papers>.
- [16] Holland, P. W., and Leinhardt, S. (1976), “Local Structure in Social Networks,” *Sociological Methodology*, 1–45.
- [17] — (1981), “An exponential family of probability distributions for directed graphs,” *Journal of the American Statistical Association*, 76, 33–65.
- [18] Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008), “Goodness of fit of social network models,” *Journal of the American Statistical Association*, 103, 248–258.
- [19] Hunter, D. R., and Handcock, M. S. (2006), “Inference in curved exponential family models for networks,” *Journal of Computational and Graphical Statistics*, 15, 565–583.
- [20] Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012), “Computational statistical methods for social network models,” *Journal of Computational and Graphical Statistics*, 21, 856–882.

- [21] Kontorovich, L., and Ramanan, K. (2008), “Concentration inequalities for dependent random variables via the martingale method,” *The Annals of Probability*, 36, 2126–2158.
- [22] Krivitsky, P. N. (2012), “Exponential-family models for valued networks,” *Electronic Journal of Statistics*, 6, 1100–1128.
- [23] Krivitsky, P. N., and Kolaczyk, E. D. (2015), “On the question of effective sample size in network modeling: An asymptotic inquiry,” *Statistical Science*, 30, 184–198.
- [24] Lazega, E., and Snijders, T. A. B. (eds.) (2016), *Multilevel Network Analysis for the Social Sciences*, Switzerland: Springer.
- [25] Lei, J., and Rinaldo, A. (2015), “Consistency of spectral clustering in stochastic block models,” *The Annals of Statistics*, 43, 215–237.
- [26] Lusher, D., Koskinen, J., and Robins, G. (2013), *Exponential Random Graph Models for Social Networks*, Cambridge, UK: Cambridge University Press.
- [27] Nowicki, K., and Snijders, T. A. B. (2001), “Estimation and prediction for stochastic blockstructures,” *Journal of the American Statistical Association*, 96, 1077–1087.
- [28] Rinaldo, A., Fienberg, S. E., and Zhou, Y. (2009), “On the geometry of discrete exponential families with application to exponential random graph models,” *Electronic Journal of Statistics*, 3, 446–484.
- [29] Rinaldo, A., Petrovic, S., and Fienberg, S. E. (2013), “Maximum likelihood estimation in network models,” *The Annals of Statistics*, 41, 1085–1110.
- [30] Rohe, K., Chatterjee, S., and Yu, B. (2011), “Spectral clustering and the high-dimensional stochastic block model,” *The Annals of Statistics*, 39, 1878–1915.
- [31] Rohe, K., Qin, T., and Fan, H. (2014), “The highest-dimensional stochastic block model with a regularized estimator,” *Statistica Sinica*, 24, 1771–1786.
- [32] Schweinberger, M. (2011), “Instability, sensitivity, and degeneracy of discrete exponential families,” *Journal of the American Statistical Association*, 106, 1361–1370.
- [33] Schweinberger, M., and Handcock, M. S. (2015), “Local dependence in random graph models: characterization, properties and statistical inference,” *Journal of the Royal Statistical Society B*, 77, 647–676.
- [34] Schweinberger, M., and Luna, P. (2015), “HERGM: Hierarchical exponential-family random graph models,” Tech. rep., Department of Statistics, Rice University.
- [35] Schweinberger, M., and Stewart, J. (2016), “Consistent  $M$ -estimation of curved exponential-family random graph models with local dependence and growing neighborhoods,” <http://arxiv.org/abs/1702.01812>.

- [36] Shalizi, C. R., and Rinaldo, A. (2013), “Consistency under sampling of exponential random graph models,” *The Annals of Statistics*, 41, 508–535.
- [37] Snijders, T. A. B. (2007), “Contribution to the discussion of Handcock, M.S., Raftery, A.E., and J.M. Tantrum, Model-based clustering for social networks,” *Journal of the Royal Statistical Society A*, 170, 322–324.
- [38] Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), “New specifications for exponential random graph models,” *Sociological Methodology*, 36, 99–153.
- [39] Stephens, M. (2000), “Dealing with label-switching in mixture models,” *Journal of the Royal Statistical Society B*, 62, 795–809.
- [40] Vu, D. Q., Hunter, D. R., and Schweinberger, M. (2013), “Model-based clustering of large networks,” *The Annals of Applied Statistics*, 7, 1010–1039.
- [41] Wasserman, S., and Pattison, P. (1996), “Logit models and logistic regression for social networks: I. An introduction to Markov graphs and  $p^*$ ,” *Psychometrika*, 61, 401–425.
- [42] Yan, T., Leng, C., and Zhu, J. (2016), “Asymptotics in directed exponential random graph models with an increasing bi-degree sequence,” *The Annals of Statistics*, 44, 31–57.

## A Proofs of auxiliary results

We prove Lemmas 1, 2, and 3 and Corollaries 1 and 2.

**PROOF OF LEMMA 1.** By assumption,  $\mathbb{E} f(\mathbf{X}) < \infty$ . We are interested in deviations of the form  $|f(\mathbf{X}) - \mathbb{E} f(\mathbf{X})| \geq t$ , where  $t > 0$ . In the following, we denote by  $\mathbb{P}$  a probability measure on  $(\mathbb{X}, \mathbb{S})$  with densities of the form (2), where  $\mathbb{S}$  is the power set of the countable set  $\mathbb{X}$ . Let  $\mathbf{X} = (\mathbf{X}_{k,l})_{k \leq l}^K$  be a sequence of edge variables, where  $\mathbf{X}_{k,k} = (X_{i,j})_{i \in \mathcal{A}_k < j \in \mathcal{A}_k}$  denotes the sequence of within-neighborhood edge variables of nodes in neighborhood  $\mathcal{A}_k$  and  $\mathbf{X}_{k,l} = (X_{i,j})_{i \in \mathcal{A}_k, j \in \mathcal{A}_l}$  denotes the sequence of between-neighborhood edge variables between nodes in neighborhoods  $\mathcal{A}_k$  and  $\mathcal{A}_l$  ( $k < l$ ). In an abuse of notation, we denote the elements of the sequence of edge variables  $\mathbf{X}$  by  $X_1, \dots, X_m$  with sample spaces  $\mathbb{X}_1, \dots, \mathbb{X}_m$ , respectively, where  $m = \binom{n}{2} \leq n^2$  is the number of edge variables. Let  $\mathbf{X}_{i:j} = (X_i, \dots, X_j)$  be a subsequence of edge variables with sample space  $\mathbb{X}_{i:j}$ , where  $i \leq j$ . By applying Theorem 1.1 of Kontorovich and Ramanan [21] to  $\|f\|_{\text{Lip}}$ -Lipschitz functions  $f : \mathbb{X} \mapsto \mathbb{R}$  defined on the countable set  $\mathbb{X}$ ,

$$\mathbb{P}(|f(\mathbf{X}) - \mathbb{E} f(\mathbf{X})| \geq t) \leq 2 \exp \left( -\frac{t^2}{2m \|\Phi\|_\infty^2 \|f\|_{\text{Lip}}^2} \right),$$

where  $\Phi$  is the  $m \times m$ -upper triangular matrix with entries

$$\phi_{i,j} = \begin{cases} \varphi_{i,j} & \text{if } i < j \\ 1 & \text{if } i = j \\ 0 & \text{if } i > j \end{cases}$$

and

$$\|\Phi\|_\infty = \max_{1 \leq i \leq m} \left| 1 + \sum_{j=i+1}^m \varphi_{i,j} \right|.$$

The coefficients  $\varphi_{i,j}$  are known as mixing coefficients and are defined by

$$\varphi_{i,j} \equiv \sup_{\substack{\mathbf{x}_{1:i-1} \in \mathbb{X}_{1:i-1} \\ (x_i, x_i^*) \in \mathbb{X}_i \times \mathbb{X}_i}} \varphi_{i,j}(\mathbf{x}_{1:i-1}, x_i, x_i^*) = \sup_{\substack{\mathbf{x}_{1:i-1} \in \mathbb{X}_{1:i-1} \\ (x_i, x_i^*) \in \mathbb{X}_i \times \mathbb{X}_i}} \|\pi_{x_i} - \pi_{x_i^*}\|_{\text{TV}},$$

where  $\|\pi_{x_i} - \pi_{x_i^*}\|_{\text{TV}}$  is the total variation distance between the distributions  $\pi_{x_i}$  and  $\pi_{x_i^*}$  given by

$$\pi_{x_i} \equiv \pi(\mathbf{x}_{j:m} \mid \mathbf{x}_{1:i-1}, x_i) = \mathbb{P}(\mathbf{X}_{j:m} = \mathbf{x}_{j:m} \mid \mathbf{X}_{1:i-1} = \mathbf{x}_{1:i-1}, X_i = x_i)$$

and

$$\pi_{x_i^*} \equiv \pi(\mathbf{x}_{j:m} \mid \mathbf{x}_{1:i-1}, x_i^*) = \mathbb{P}(\mathbf{X}_{j:m} = \mathbf{x}_{j:m} \mid \mathbf{X}_{1:i-1} = \mathbf{x}_{1:i-1}, X_i = x_i^*).$$

Since the support of  $\pi_{x_i}$  and  $\pi_{x_i^*}$  is countable,

$$\|\pi_{x_i} - \pi_{x_i^*}\|_{\text{TV}} = \frac{1}{2} \sum_{\mathbf{x}_{j:m} \in \mathbb{X}_{j:m}} |\pi(\mathbf{x}_{j:m} \mid \mathbf{x}_{1:i-1}, x_i) - \pi(\mathbf{x}_{j:m} \mid \mathbf{x}_{1:i-1}, x_i^*)|.$$

An upper bound on  $\|\Phi\|_\infty$  can be obtained by bounding the mixing coefficients  $\varphi_{i,j}$  as follows. Consider any pair of edge variables  $X_i$  and  $X_j$ . If  $X_i$  and  $X_j$  involve nodes in more than one neighborhood, the mixing coefficient  $\varphi_{i,j}$  vanishes by the local dependence induced by exponential families with local dependence. If the pair of nodes corresponding to  $X_i$  and the pair of nodes corresponding to  $X_j$  belong to the same neighborhood, the mixing coefficient  $\varphi_{i,j}$  can be bounded as follows:

$$\begin{aligned}\varphi_{i,j}(\mathbf{x}_{1:i-1}, x_i, x_i^*) &= \frac{1}{2} \sum_{\mathbf{x}_{j:m} \in \mathbb{X}_{j:m}} |\pi(\mathbf{x}_{j:m} \mid \mathbf{x}_{1:i-1}, x_i) - \pi(\mathbf{x}_{j:m} \mid \mathbf{x}_{1:i-1}, x_i^*)| \\ &\leq \frac{1}{2} \sum_{\mathbf{x}_{j:m} \in \mathbb{X}_{j:m}} \pi(\mathbf{x}_{j:m} \mid \mathbf{x}_{1:i-1}, x_i) + \frac{1}{2} \sum_{\mathbf{x}_{j:m} \in \mathbb{X}_{j:m}} \pi(\mathbf{x}_{j:m} \mid \mathbf{x}_{1:i-1}, x_i^*) = 1,\end{aligned}$$

because  $\pi_{x_i}$  and  $\pi_{x_i^*}$  are conditional probability mass functions with countable support  $\mathbb{X}_{j:m}$ . We note that the upper bound is not sharp, but it has the advantage that it covers a wide range of dependencies within neighborhoods. As a result,

$$\|\Phi\|_\infty = \max_{1 \leq i \leq m} \left| 1 + \sum_{j=i+1}^m \varphi_{i,j} \right| \leq \binom{\|\mathcal{A}\|_\infty}{2},$$

because each edge variable  $X_i$  can depend on at most  $\binom{\|\mathcal{A}\|_\infty}{2}$  edge variables corresponding to pairs of nodes belonging to the same pair of neighborhoods. Therefore, there exists  $C > 0$  such that, for all  $K > 0$  and all  $t > 0$ ,

$$\mathbb{P}(|f(\mathbf{X}) - \mathbb{E} f(\mathbf{X})| \geq t) \leq 2 \exp \left( -\frac{t^2}{C n^2 \|\mathcal{A}\|_\infty^4 \|f\|_{\text{Lip}}^2} \right),$$

where  $\|\mathcal{A}\|_\infty > 0$  and  $\|f\|_{\text{Lip}} > 0$  by assumption.

**PROOF OF LEMMA 2.** Since the data-generating natural parameter vector  $\boldsymbol{\eta}^* \in \Xi \subseteq \text{int}(\mathbb{N})$  is in the interior  $\text{int}(\mathbb{N})$  of the natural parameter space  $\mathbb{N}$ , the expectation  $\mathbb{E} s(\mathbf{X})$  exists [7, Theorem 2.2, pp. 34–35] and so does the expectation  $\mathbb{E} \ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{X})) = \ell(\boldsymbol{\theta}, \mathbf{z}; \mathbb{E} s(\mathbf{X}))$ . We want to bound

$$\mathbb{P}(\mathbf{X} \in \mathbb{X} \setminus \mathbb{X}(\alpha)) = \mathbb{P}(|\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; s(\mathbf{X})) - \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)| \geq \alpha u(n)),$$

where

$$u(n) = |\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|.$$

Bounding the probability of deviations of the form  $|\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; s(\mathbf{X})) - \ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)|$  is equivalent to bounding the probability of deviations of the form  $|f(\mathbf{X}) - \mathbb{E} f(\mathbf{X})|$ , where

$$f(\mathbf{X}) = \langle \boldsymbol{\eta}(\boldsymbol{\theta}^*, \mathbf{z}^*), s(\mathbf{X}) \rangle, \quad \mathbb{E} f(\mathbf{X}) = \langle \boldsymbol{\eta}(\boldsymbol{\theta}^*, \mathbf{z}^*), \boldsymbol{\mu}^* \rangle.$$

We note that  $f : \mathbb{X} \mapsto \mathbb{R}$  is considered as a function of  $\mathbf{X}$  for fixed  $(\boldsymbol{\theta}^*, \mathbf{z}^*) \in \boldsymbol{\Theta}_0 \times \mathbb{Z}_0$  and that  $\psi(\boldsymbol{\eta}(\boldsymbol{\theta}^*, \mathbf{z}^*))$  cancels. Observe that by condition [C.4] there exist  $A_2 > 0$  and  $n_0 > 0$

such that, for all  $n > n_0$ , the Lipschitz coefficient of  $f(\mathbf{X})$  satisfies  $\|f\|_{\text{Lip}} \leq A_2 L$ . Thus, by applying Lemma 1 to deviations of size  $t = \alpha u(n)$ , there exist  $C_0 > 0$  and  $n_0 > 0$  such that, for all  $n > n_0$ ,

$$\mathbb{P}(\mathbf{X} \in \mathbb{X} \setminus \mathbb{X}(\alpha)) \leq 2 \exp \left( -\frac{\alpha^2 u(n)^2}{C_0 n^2 \|\mathcal{A}\|_\infty^4 L^2} \right).$$

By assumption (4) of Proposition 1, there exists, for all  $C_1 > 0$ , however large,  $n_1 > 0$  such that, for all  $n > n_1$ ,

$$u(n) \geq C_1 n^{3/2} \|\mathcal{A}\|_\infty^2 L \sqrt{\log n}.$$

Therefore, there exists  $C > 0$  such that, for all  $n > \max(n_0, n_1)$ ,

$$\mathbb{P}(\mathbf{X} \in \mathbb{X} \setminus \mathbb{X}(\alpha)) \leq 2 \exp(-\alpha^2 C n \log n).$$

**PROOF OF LEMMA 3.** In the following, we confine attention to  $\mathbf{x} \in \mathbb{X}(\alpha)$ , because we are interested in the existence of the restricted maximum likelihood estimator  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  in the event  $\mathbb{X}(\alpha)$ . For any  $\mathbf{x} \in \mathbb{X}(\alpha)$  and any  $\mathbf{z} \in \mathbb{Z}_0$ , let

$$\hat{\boldsymbol{\theta}}(\mathbf{z}) = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{x})).$$

Observe that, for any  $\mathbf{x} \in \mathbb{X}(\alpha)$  and any  $\mathbf{z} \in \mathbb{Z}_0$ , the loglikelihood function  $\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{x}))$  is upper semicontinuous on  $\boldsymbol{\Theta}_0$  by condition [C.2]. In addition, by condition [C.5] there exist  $A, B, C > 0$  such that the  $\dim(\boldsymbol{\theta}) \leq A n$ -dimensional parameter space  $\boldsymbol{\Theta}_0$  can be covered by  $\exp(Cn)$  closed balls with centers  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$  and radius  $B > 0$ . As a result, for any  $\mathbf{x} \in \mathbb{X}(\alpha)$  and any  $\mathbf{z} \in \mathbb{Z}_0$ ,  $\ell(\boldsymbol{\theta}, \mathbf{z}; s(\mathbf{x}))$  assumes a maximum on  $\boldsymbol{\Theta}_0$  and hence the maximizer  $\hat{\boldsymbol{\theta}}_l(\mathbf{z})$  exists and is unique by condition [C.1] and the assumption that the exponential family is minimal, which can be assumed without loss [7, Theorem 1.9, p. 13]. Since, for any  $\mathbf{z} \in \mathbb{Z}_0$ ,  $\hat{\boldsymbol{\theta}}(\mathbf{z})$  exists, so does  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$ .

Last, but not least, since  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  exists for all  $\mathbf{x} \in \mathbb{X}(\alpha)$ ,  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  exists with at least probability  $\mathbb{P}(\mathbf{X} \in \mathbb{X}(\alpha))$ . By Lemma 2, there exist  $C_0 > 0$  and  $n_0 > 0$  such that, for all  $n > n_0$ ,

$$\mathbb{P}(\mathbf{X} \in \mathbb{X}(\alpha)) \geq 1 - 2 \exp(-\alpha^2 C_0 n \log n).$$

Therefore, for all  $n > n_0$ ,  $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$  exists with at least probability  $1 - 2 \exp(-\alpha^2 C_0 n \log n)$ .

**PROOF OF COROLLARY 1.** To show that conditions [C.1]—[C.4] are satisfied, note that  $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \Xi$  is separable in the sense that  $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}) = \mathbf{A}(\mathbf{z}) \mathbf{b}(\boldsymbol{\theta})$  and hence  $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})$  can be reduced to  $\boldsymbol{\eta}(\boldsymbol{\theta})$  by absorbing  $\mathbf{A}(\mathbf{z})$  into the sufficient statistics vector. In addition, since the exponential family is canonical,  $\boldsymbol{\eta}(\boldsymbol{\theta})$  can be reduced to  $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ . Condition [C.1] is satisfied because  $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ . Condition [C.2] follows from  $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$  and the upper semicontinuity of canonical exponential-family loglikelihood functions [7, Lemma 5.3, p. 146]. To show that condition [C.3] holds, observe that

$$|\langle \boldsymbol{\eta}(\boldsymbol{\theta}_1, \mathbf{z}) - \boldsymbol{\eta}(\boldsymbol{\theta}_2, \mathbf{z}), \boldsymbol{\mu} \rangle| = |\langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \boldsymbol{\mu}(\mathbf{z}) \rangle|,$$



where  $\boldsymbol{\mu}(\mathbf{z}) = \mathbf{A}(\mathbf{z})^\top \boldsymbol{\mu}$  ( $\boldsymbol{\mu} \in \mathbb{M}(\alpha)$ ). We can therefore write

$$\begin{aligned}
|\langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \boldsymbol{\mu}(\mathbf{z}) \rangle| &= \sum_{k \leq l}^K |\langle \boldsymbol{\theta}_{1,k,l} - \boldsymbol{\theta}_{2,k,l}, \boldsymbol{\mu}_{k,l}(\mathbf{z}) \rangle| \\
&\leq \sum_{k \leq l}^K \|\boldsymbol{\theta}_{1,k,l} - \boldsymbol{\theta}_{2,k,l}\|_1 \|\boldsymbol{\mu}_{k,l}(\mathbf{z})\|_\infty \\
&\leq \sum_{k \leq l}^K \sqrt{\dim(\boldsymbol{\theta}_{k,l})} \|\boldsymbol{\theta}_{1,k,l} - \boldsymbol{\theta}_{2,k,l}\|_2 \|\boldsymbol{\mu}_{k,l}(\mathbf{z})\|_\infty.
\end{aligned}$$

Since the parameter vectors  $\boldsymbol{\theta}_{k,l}$  are finite-dimensional, the parameter space  $\boldsymbol{\Theta}_0$  is compact, and the random graph is dense in the sense that  $|\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)| = C_0 \binom{n}{2}$  ( $C_0 > 0$ ), condition [C.3] is satisfied as long as  $\|\boldsymbol{\mu}_{k,l}(\mathbf{z})\|_\infty \leq C_1 L_k(\mathbf{z}) L_l(\mathbf{z})$  ( $C_1 > 0$ ) for all  $\boldsymbol{\mu} \in \mathbb{M}(\alpha)$  and all  $\mathbf{z} \in \mathbb{Z}_0$ . The same argument shows that

$$\begin{aligned}
|\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), s(\mathbf{x}_1) - s(\mathbf{x}_2) \rangle| &= |\langle \boldsymbol{\theta}, s(\mathbf{x}_1, \mathbf{z}) - s(\mathbf{x}_2, \mathbf{z}) \rangle| \\
&= \sum_{k \leq l}^K |\langle \boldsymbol{\theta}_{k,l}, s_{k,l}(\mathbf{x}_1, \mathbf{z}) - s_{k,l}(\mathbf{x}_2, \mathbf{z}) \rangle| \\
&\leq \sum_{k \leq l}^K \sqrt{\dim(\boldsymbol{\theta}_{k,l})} \|\boldsymbol{\theta}_{k,l}\|_2 \|s_{k,l}(\mathbf{x}_1, \mathbf{z}) - s_{k,l}(\mathbf{x}_2, \mathbf{z})\|_\infty.
\end{aligned}$$

As a result, condition [C.4] is satisfied as long as  $\sum_{k \leq l}^K \|s_{k,l}(\mathbf{x}_1, \mathbf{z}) - s_{k,l}(\mathbf{x}_2, \mathbf{z})\|_\infty \leq C_2 d(\mathbf{x}_1, \mathbf{x}_2) L(\mathbf{z})$  for all  $(\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{X} \times \mathbb{X}$  and all  $\mathbf{z} \in \mathbb{Z}_0$ .

**PROOF OF COROLLARY 2.** To streamline the presentation, we assume the following:

- We take advantage of the fact that  $\boldsymbol{\eta} : \boldsymbol{\Theta} \times \mathbb{Z} \mapsto \Xi$  is separable in the sense that  $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}) = \mathbf{A}(\mathbf{z}) \mathbf{b}(\boldsymbol{\theta})$  and reduce  $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z})$  to  $\boldsymbol{\eta}(\boldsymbol{\theta})$  by absorbing  $\mathbf{A}(\mathbf{z})$  into the sufficient statistics vector.
- Since, under the curved exponential-family random graph model (6) described in Section 3.2, between- and within-neighborhood edge terms cannot violate conditions [C.1]–[C.4], we assume that there is a single neighborhood without edge terms but with geometrically weighted model terms of the form (6), so that we can write

$$\begin{aligned}
\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), \boldsymbol{\mu} \rangle &= \langle \boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{\mu}(\mathbf{z}) \rangle = \sum_{t=1}^T \eta_t(\boldsymbol{\theta}) \mu_t(\mathbf{z}) \\
\langle \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{z}), s(\mathbf{x}) \rangle &= \langle \boldsymbol{\eta}(\boldsymbol{\theta}), s(\mathbf{x}, \mathbf{z}) \rangle = \sum_{t=1}^T \eta_t(\boldsymbol{\theta}) s_t(\mathbf{x}, \mathbf{z}),
\end{aligned}$$

where  $\boldsymbol{\mu}(\mathbf{z}) = \mathbf{A}(\mathbf{z})^\top \boldsymbol{\mu}$  ( $\boldsymbol{\mu} \in \mathbb{M}(\alpha)$ ) and  $s(\mathbf{x}, \mathbf{z}) = \mathbf{A}(\mathbf{z})^\top s(\mathbf{x})$  ( $s(\mathbf{x}) \in \mathbb{M}(\alpha)$ ). Throughout, we drop the subscript  $k$ —which indexes neighborhoods—from all neighborhood-dependent quantities, because there is a single neighborhood.

- We assume that the parameter  $\theta_1$  of the within-neighborhood edge term and the base parameter  $\theta_2$  of the within-neighborhood geometrically weighted model term are given by  $\theta_1 = 0$  and  $\theta_2 = 1$ , respectively, and drop the subscript of  $\theta_3$ , i.e., we write  $\theta$  rather than  $\theta_3$ .

The extension to more than one neighborhood and  $\theta_1 \in \mathbb{R}$  is straightforward. The extension to  $\theta_2 \in \mathbb{R}$  can be proved along the lines of Schweinberger and Stewart [35].

Under the assumptions outlined above, the coordinates  $\eta_t(\theta)$  of the single within-neighborhood natural parameter vector  $\boldsymbol{\eta}(\theta)$  can be written as

$$\eta_t(\theta) = \theta \left[ 1 - \left( 1 - \frac{1}{\theta} \right)^t \right] = \theta - \theta \beta(\theta)^t, \quad t = 1, \dots, T, \quad (10)$$

where

$$\beta(\theta) = 1 - \frac{1}{\theta}.$$

The parameter space  $\Theta$  is given by

$$\Theta = \left\{ \theta \in \mathbb{R} : \frac{1}{2} < \theta < B, \psi(\boldsymbol{\eta}(\theta, \mathbf{z})) < \infty \right\}, \quad B > \frac{1}{2}.$$

A helpful observation is that the coordinates  $\eta_t(\theta)$  of  $\boldsymbol{\eta}(\theta)$  are continuously differentiable on  $(1/2, B)$  with derivatives

$$\nabla_\theta \eta_t(\theta) = 1 - \beta(\theta)^t - \frac{t}{\theta} \beta(\theta)^{t-1}, \quad \theta \in (1/2, B).$$

We check conditions [C.1]—[C.4] one by one.

Condition [C.1]. To show that the map  $\boldsymbol{\eta} : \Theta \mapsto \Xi$  is one-to-one on  $\Theta$ , we show that at least one coordinate of  $\boldsymbol{\eta}(\theta + \delta)$  must deviate from  $\boldsymbol{\eta}(\theta)$  for all  $\theta \in (1/2, B)$  and all  $\delta > 0$ . To do so, note that  $\boldsymbol{\eta}(\theta)$  has at least two coordinates, denoted by  $\eta_1(\theta)$  and  $\eta_2(\theta)$ , because  $T \geq 2$  by assumption. The first coordinate  $\eta_1(\theta)$  of  $\boldsymbol{\eta}(\theta)$  is constant on  $(1/2, B)$ :

$$\eta_1(\theta) = 1, \quad \theta \in (1/2, B).$$

The second coordinate  $\eta_2(\theta)$  of  $\boldsymbol{\eta}(\theta)$  is continuously differentiable on  $(1/2, B)$  with derivative

$$\nabla_\theta \eta_2(\theta) = 1 - \beta(\theta)^2 - \frac{2}{\theta} \beta(\theta) = \frac{1}{\theta^2} > 0, \quad \theta \in (1/2, B).$$

By the mean-value theorem,

$$\eta_2(\theta + \delta) - \eta_2(\theta) \geq \frac{\delta}{(\theta + \delta)^2} > 0, \quad \theta \in (1/2, B), \quad \delta > 0.$$

Thus,  $\eta_2(\theta)$  is strictly increasing on  $(1/2, B)$  and at least one coordinate of  $\boldsymbol{\eta}(\theta + \delta)$  must deviate from  $\boldsymbol{\eta}(\theta)$  for all  $\theta \in (1/2, B)$  and all  $\delta > 0$ . As a result, the map  $\boldsymbol{\eta} : \Theta \mapsto \Xi$  is one-to-one and continuous on  $\Theta$ . Thus condition [C.1] is satisfied.

Condition [C.2]. Condition [C.2] follows from the continuity of  $\boldsymbol{\eta} : \Theta \mapsto \Xi$  and the upper semicontinuity of exponential-family loglikelihood functions [7, Lemma 5.3, p. 146].

Condition [C.3]. Choose any  $\theta \in \Theta$  and  $\theta' \in \Theta$  and let  $\boldsymbol{\mu}(\mathbf{z}) = \mathbf{A}(\mathbf{z})^\top \boldsymbol{\mu}$  ( $\boldsymbol{\mu} \in \mathbb{M}(\alpha)$ ). By the triangle inequality, we obtain, for all  $\theta \in \Theta$  and  $\theta' \in \Theta$  and all  $\boldsymbol{\mu} \in \mathbb{M}(\alpha)$ ,

$$\begin{aligned} |\langle \boldsymbol{\eta}(\theta') - \boldsymbol{\eta}(\theta), \boldsymbol{\mu}(\mathbf{z}) \rangle| &= \left| \sum_{t=1}^T [\eta_t(\theta') - \eta_t(\theta)] \mu_t(\mathbf{z}) \right| \\ &\leq \sum_{t=1}^T |\eta_t(\theta') - \eta_t(\theta)| |\mu_t(\mathbf{z})|. \end{aligned} \tag{11}$$

It can be shown [35] that there exists  $C > 2$  such that, for all  $\theta \in \Theta$  and all  $t \in \{1, 2, \dots\}$ ,

$$|\nabla_\theta \eta_t(\theta)| \leq \max(3, C), \quad t \in \{1, 2, \dots\},$$

which, by the mean-value theorem, implies that

$$|\eta_t(\theta') - \eta_t(\theta)| \leq |\theta' - \theta| \max(3, C), \quad t \in \{1, 2, \dots\}.$$

Using (11) along with condition [C.3\*\*] shows that there exist  $C_1 > 0$  and  $n_1 \geq 1$  such that, for all  $n > n_1$ ,

$$\begin{aligned} |\langle \boldsymbol{\eta}(\theta') - \boldsymbol{\eta}(\theta), \boldsymbol{\mu}(\mathbf{z}) \rangle| &\leq \sum_{t=1}^T |\eta_t(\theta') - \eta_t(\theta)| |\mu_t(\mathbf{z})| \\ &\leq |\theta' - \theta| \max(3, C) \sum_{t=1}^T |\mu_t(\mathbf{z})| \leq C_1 \|\theta' - \theta\|_2 \binom{n}{2}. \end{aligned}$$

Hence condition [C.3] is satisfied, because  $|\ell(\boldsymbol{\theta}^*, \mathbf{z}^*; \boldsymbol{\mu}^*)| = C \binom{n}{2}$  ( $C > 0$ ) in dense random graphs and because we assume that there is a single neighborhood.

Condition [C.4]. Using  $|\beta(\theta)| < 1$  for all  $\theta \in \Theta$ ,

$$|\eta_t(\theta)| \leq |\theta| + |\theta| |\beta(\theta)|^t \leq 2B, \quad \theta \in \Theta.$$

By condition [C.4\*\*], there exist  $C_2 > 0$  and  $n_2 \geq 1$  such that, for all  $n > n_2$ ,

$$\begin{aligned} |\langle \boldsymbol{\eta}(\theta), s(\mathbf{x}_1, \mathbf{z}) - s(\mathbf{x}_2, \mathbf{z}) \rangle| &= \left| \sum_{t=1}^T \eta_t(\theta) [s_t(\mathbf{x}_1, \mathbf{z}) - s_t(\mathbf{x}_2, \mathbf{z})] \right| \\ &\leq 2B \left| \sum_{t=1}^T s_t(\mathbf{x}_1, \mathbf{z}) - \sum_{t=1}^T s_t(\mathbf{x}_2, \mathbf{z}) \right| \leq C_2 d(\mathbf{x}_1, \mathbf{x}_2) L(\mathbf{z}). \end{aligned}$$

Thus condition [C.4] is satisfied.